TU Clausthal

Clausthal University of Technology

ICOLE 2008, Lessach, Austria

Jacek Blazewicz, Klaus Ecker, Barbara Hammer (Eds.)

IfI Technical Report Series **IfI-08-06**

Department of Informatics
Clausthal University of Technology

Impressum

Publisher: Institut für Informatik, Technische Universität Clausthal
Julius-Albert Str. 4, 38678 Clausthal-Zellerfeld, Germany
Editor of the series: Jürgen Dix
Technical editor: Wojciech Jamroga
Contact: wjamroga@in.tu-clausthal.de
URL: http://www.in.tu-clausthal.de/forschung/technical-reports/
ISSN: 1860-8477

The IfI Review Board

ICOLE - 2008
German - Polish Workshop on Computational Biology, Scheduling and Machine Learning
Lessach, 9.06. - 13.06.2008

Jacek Blazewicz, Klaus Ecker, Barbara Hammer (Eds.)

Managing Editors: Bassam Mokbel, Wibke Börger, Alexander Hasenfuss

Contents

Preview

Jacek Blazewicz[1,2]***, Klaus Ecker***[3,4]***, Barbara Hammer***[5,6]

From June 9 to 13 2008, seven scientists from the Clausthal University of Technology and 13 scientists from the Poznan University of Technology met in Daublebsky's wonderful house in Lessach, Austria, to present topics of their current research and share their ideas. The actual report collects the abstracts of altogether 17 presentations from the following areas from Computer Science:

- Large data sets: One group of presentations concentrated around structuring and visualizing large data sets and discussed theoretical basics and concepts as well as applications in large text systems, music, time series, and cognitive sciences.
- DNA and RNA: Another group dealt with new approaches to sequencing and assembling DNA sequences, graph models for hybridization of DNA chains, and computational methods for elucidating three-dimensional RNA structures.
- Assembly line balancing: Two presentations discussed problems of workers' assignment for balancing assembly lines in vehicle scheduling.
- Quantum computing: A report on quantum computing informed about the status and some aspects of quantum computing.

There were also three reports about European projects:

- CompuVac: The European project COPMUVAC aims on setting up a standardized approach for the rational development of genetic vaccines, with particular application to the development of vaccines against the hepatitis C virus.

[1]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[2]E-mail: jblazewicz@cs.put.poznan.pl
[3]Center for Intelligent, Distributed and Dependable Systems, Ohio University, Athens, USA
[4]E-mail: ecker@ohio.edu
[5]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[6]E-mail: hammer@in.tu-clausthal.de

- MetaFunctions: the main goal of the EU project METAFUNCTIONS is creating a bioinformatic system to detect and assign functions of habitat specific gene patterns. Poseidon, being a part of the system, classifies scientific documents concerning marine metagenomics and extracts the relevant data.
- Simulation of Crisis Management: Report on the EU project "Simulation of Crisis Management Activities" (SICMA), and supports from resource constraint scheduling theory.

The seminar continued a series of similar events, organized during the last years by both institutions. As always, this workshop was a great success due to its internationality, the friendly and cooperative atmosphere during the seminar, and the excellent possibilities offered by the house.

Lessach, June 14, 2008
Jacek Blazewicz, Klaus Ecker, Barbara Hammer

Some Facts About Recurrent Neural Systems

Barbara Hammer[1,2]

1 Introduction

Artificial neural networks (FNNs) constitute a particularly successful machine learning technique which applications range from industrial tasks up to simulations of biological neural networks. Recurrent neural networks (RNNs) extend FNNs by cyclic connections, such that they can incorporate temporal, spatial, or causal dependencies in a natural way. Such dependencies occur e.g. in robotics, system identification and control, bioinformatics, medical and biomedical data such as EEG and EKG, sensor streams in technical applications, natural speech processing, analysis of text and web documents, etc. Fruther, spatiotemporal signals and feedback connections are ubiquitous when considering biological neural networks of the human brain. Thus, RNNs carry the promise of efficient biologically plausible signal processing models optimally suited for a wide area of industrial applications on the one hand and an explanation of cognitive phenomena of the human brain on the other hand.

However, simple feedforward networks without recurrent connections are still the preferred model in industrial or scientific applications although they neglect structural aspects of data and require an often time-consuming feature-encoding of all information. This is mainly due to one reason – unlike FNNs, training RNNs by means of virtually all currently known mechanisms faces severe problems: backpropagation for RNNs suffers from numerical barriers, a formal learning theory of RNNs in the classical sense of PAC learning does hardly exist, RNNs easily show complex chaotic behavior which is complicated to manage, and the way how humans use recurrence to cope with language, complex symbols, or logical inference is only partially understood.

In recent years, several new fundamental paradigms connected to RNNs have been developed which allow new insights into potential information processing with RNNs and open the way to new efficient training algorithms. Thereby, various results have been achieved addressing different goals, including the explanation of 'recurrent' phenomena observed in humans and the development of corresponding biologically plau-

[1]Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[2]E-mail: hammer@in.tu-clausthal.de

sible models, the design of efficient recurrent training algorithms beyond numerically instable gradient based techniques, the theoretical understanding of the capacity of recurrent models, its connections to high-level structures and symbolic paradigms, and the design of corresponding systems. The aim of this contribution is to give an overview about the capacity of recurrent systems in comparison to classical symbolic formalisms.

2 Supervised recurrent systems

Supervised recurrent systems aim at learning an input-output relationship from given data, such as time series prediction, or transduction. Classical training algorithms for discrete time recurrent networks include backpropagation through time, real time recurrent learning, Kalman filtering, and variations thereof. Unfortunately, these mechanisms rely on gradient information which cannot capture long term dependencies in a reliable way [1]. In recent years, a trend towards formalisms which do not require (complex) training of the recurrent part of neural systems has been proposed. This includes fractal prediction machines, echo state networks, and liquid state machines as most prominent examples [5, 6, 8]. although these methods restrict the recurrent part considerably, they show surprisingly good results in a variety of applications. Besides, they offer a striking biological plausibility. Now the question occurs what is lost by fixing the recurrent part of RNNs in advance independent of the learning task.

The capacity of classical RNNs is well established and can be set into relation to the classical Chomsky hierarchy, ranging from non-uniform Boolean circuits for general RNNs which possess super-Turing capability to finite memory models if the computation is affected by Gaussian noise, see e.g. [2] for an overview. Obviously, general Turing machines cannot be learned from examples, neither do alternative formalisms if they display sufficient complexity. We argue that, for learning, RNNs with fixed recurrent part are not far from RNNs which are trained in a standard way, providing insight into the architectural bias of RNN learning. It can be shown that RNNs with small weights (the standard way of initializing networks) are equivalent to finite memory machines [4]. This puts RNNs close to echo state networks, for example, in the first steps of training. Interestingly, this restriction of the capacity of RNNs has benefitial effects on the learnability, since for small weight RNNs, unlike general RNNs, distribution independent generalization bounds can be derived (see [2]).

3 Unsupervised recurrent systems

Unsupervised recurrent networks aim at a meaningful representation of potentially high-dimensional signals with temporal or causal context, such as a display of time series events. Unlike supervised systems, no explicit teacher information is available and the overall goal is to extract the most important information from the data such as cluster structure and global topology for data inspection. For feedforward systems, several well-established data mining tools have been proposed such as the self-organizing map

(SOM) or neural gas. Interestingly, a variety of different extensions towards temporal data has been proposed independently, including the temporal Kohonen map and the recurrent SOM as early models. These methods incorporate a leaky integration into the computation of the winner, such that robust temporal averaging is achieved. However, the methods are rather restricted with respect to their capacity. Several more powerful methods have been proposed in recent years which include an explicit representation of the temporal context, such as recursive SOM, feedback SOM, SOM for structured data, and merge SOM. The models can be represented by one set of global dynamical equations and they differ only in the choice of context, as shown e.g. in [3].

Since no teacher information is available and the models are trained in a purely unsupervised fashion, usually by some form of Hebbian learning, it is of uttermost importance to understand the capacity of these models and the representation bias of learning. This is a key ingredient for a valid interpretation of the results found by unsupervised training. We present a few results along this line for different models, in particular a formal characterization of their capacity in terms of classical mechanisms of the Chomsky hierarchy and a formal characterization of the contexts which evolve during training. Interestingly, depending on the choice of the context, different capacities can be observed, ranging from finite memory models for leaky integration based approaches to finite state automata and beyond for merge SOM and recursive SOM [3, 7].

References

[1] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 1994.

[2] B. Hammer. Learning with recurrent neural networks, Springer, 2000.

[3] B.Hammer, A.Micheli, A.Sperduti, and M.Strickert. Recursive self-organizing network models, Neural Networks, 2004.

[4] B. Hammer and P. Tino. Recurrent neural networks with small weights implement definite memory machines. Neural Computation, 2003.

[5] H. Jaeger and H. Haas. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. Science, 2004.

[6] W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Computation, 2002.

[7] M.Strickert and B.Hammer. Merge SOM for temporal data. Neurocomputing, 2005.

[8] P. Tino and G. Dorffner. Predicting the future from fractal representations of the past, Machine Learning, 2001.

TU Clausthal
ICOLE 2008, LESSACH, AUSTRIA

Time Series Clustering by Recurrent SOMs

Andrej Gisbrecht[1,2]

1 Introduction

Clustering constitutes a standard method to extract typical characteristics of given data. For many application areas such as robotics, medical signals, speech, etc. data possesses a temporal structure which should be taken into account by appropriate mechanisms. Unfortunatly as shown in [3], a popular method for time series clustering, namely sliding window clustering, is meaningless in the sense that results do not reveal information about the given data at hand rather than they display general statistical properties which are present in any time series. This article gives an overview about several methods to extract motifs from time series by means of recurrent dynamics and it investigates if they are meaningful or not.

2 Recurrent self-organizing maps

The self-organizing map (SOM) constitutes a popular clustering and data visualization tool which adapts a lattice of neurons to a given data set in a topology preserving manner. SOM can be directly used for time series by means of sliding windows, however, this will yield meaningless results as shown in [3]. In recent years, a variety of extensions of SOM towards time series clustering by means of a recurrent dynamics has been proposed [2]. The main idea is to use a second layer SOM to learn the temporal context of each point. Hence neuron ij in a rectangular lattice is equipped with a weight w_{ij} and a context representation c_{ij}. The winner neuron is then determined through the distance:

$$d_{ij}(P_t) = \alpha \cdot distance(w_{ij}, P_t) + (1-\alpha) \cdot distance(c_{ij}, C_t)$$

where P_t represents the current entry of the time series, C_t the temporal context, which is specified below, and $\alpha \in (0,1)$ determines the weighting of the curent en-

[1]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[2]E-mail: andrej.gisbrecht@tu-clausthal.de

try and the context influence. Different methods to represented the context C_t are of interest in this contribution.

- **Merge SOM (MSOM):** Merge SOM has been proposed in [4].

  The weight of the winner neuron is merged with its context and the result is used as the next context C_t

- **SOM for structured data (SOMSD):** SOM for structured data has been proposed for the unsupervised processing of tree structures [1]. Restricting to linear dependencies only, time series arise. For SOMSD, the position of the winner neuron in the SOM is used as the temporal context.

- **Recursive SOM (RecSOM):** The recursive SOM has been proposed in [6]. The distance from the current point to all neurons in the SOM serves as the temporal context for the next step: $C_{t+1} = (e^{-d_{11}(P_t)}, ..., e^{-d_{nm}(P_t)})$

3 Evaluation

After the SOM is trained, we can extract the strings that are characteristic for the time series. For the extraction we can use two methods.

- **Forward processing:** We start with a neuron and calculate the context that would be generated if the weight of this neuron would be a point in a time series. Then we look for the neuron with minimal distance to this context. The weight of this neuron is the next point in the time series. Then we repeat this procedure. Thereby, stings are build from the begining to the end of the time series.

- **Backward processing:** In the second method we directly find a predecessor neuron by analyzing the context of this neuron. In the case of SOMSD the context of a neuron directly indicates the position of its predecessor. For the recursive SOM we can take the position of the largest entry of the context as predecessor. For the merge SOM we calculate the merged value of weight and context for each neuron. The one that is most similar to the considered context is its predecessor.

Now as the strings are extracted we should test whether they are meaningfull. In this way we know whether the algorithm is able to learn the time series and provide information about it. We use the definition of meaninglessness from [3]: If $A = (a_1, ..., a_n)$ and $B = (b_1, ..., b_n)$ are sets of extracted strings, then the distance between this clusterings is defined as:

$$distance(A, B) = \sum_{i=1}^{k} min(dist(a_i, b_i)), 1 \leq j \leq k$$

Let X contain 3 restarts of SOM on time series and Y contain 3 restarts of SOM on a random walk, then the meaninglessness is defined as:

$$meaninglessness(X,Y) = \frac{within\_set\_X\_distance}{between\_set\_X\_and\_Y\_distance}$$

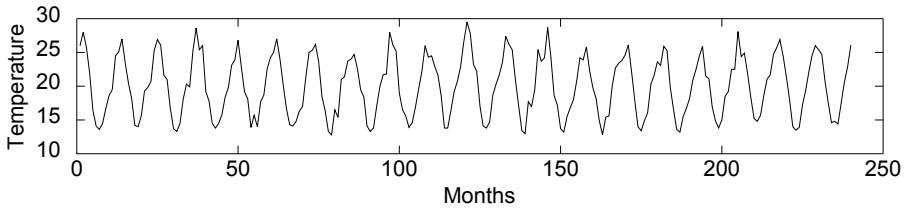


Figure 1: Mean maximum temperatures in Melbourne from 1971 to 1990

For a concrete evaluation, the time series displaying the mean maximum temperatures in Melbourne from 1971 to 1990 from [5] was used. The measurement was repeated ten times with different random walks. In each run strings with length raning from 2 to 12 were extracted.

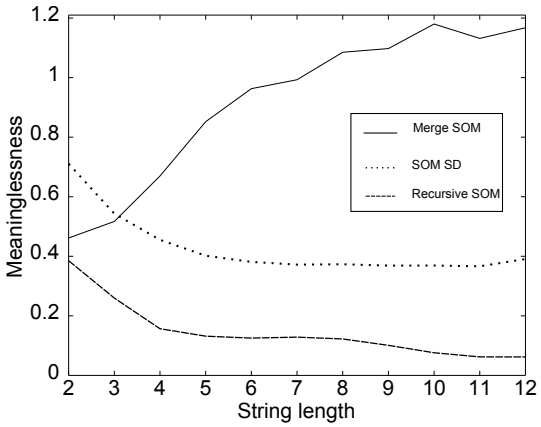


Figure 2: Mean maximum temperatures in Melbourne from 1971 to 1990

As we can see the merge SOM is quite bad for this time series. The strings that are produced from the given time series display nearly the same variance as the ones that are extracted from a random walk. SOMSD is better, but still not really good. The recursive SOM has a very small coefficient, so we can say that this is a good method for clustering this time series. This fact is not surprising since merge SOM saves very little context and the recursive SOM saves the most.

As a second experiment, we use recursive SOM for time series prediction. This is done in the following way: after processing a time series, a context is available. Then we look for the neuron that best matches this context. The weight of this neuron is

predicted as next point of the time series. If we use this method on simple time series like repeating $1-2-3-1-...$, the prediction is perfect for all three methods recSOM, SOMSD, and MSOM. On real time series the methods produce a prediction error. We compare the result with a standard model, namely the ARMA-Model. In principle, ARMA consist of two models: AR - Auto Regression and MA - Moving Average. MA simulates the white noise and can be understood as the expectation value. AR calculates the linear weighted previous values of time series and is a perturbation that depends on previous values.

Again we use the time series displaying temperatures in Melbourne. Both recursive SOM and ARMA are learned with the first half of the time series and then make one-step predictions on the second half. The prediction errors of both methods are similar. We can see that ARMA aproximates the time series through sinusoidal curves. Recursive SOM instead tries to meet the peaks and makes bigger mistakes, but seems to have a more similar overall shape to the given curve.

References

[1] M. Hagenbuchner, A. Sperduti, and A. Tsoi. A self-organizing map for adaptive processing of structured data. IEEE Transactions on Neural Networks, 14:491–505, 2003.

[2] B.Hammer, A.Micheli, A.Sperduti, M.Strickert, Recursive self-organizing network models, Neural Networks 17(8-9), 1061-1086, 2004.

[3] E. Keogh, J. Lin, W. Truppel,

Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research,

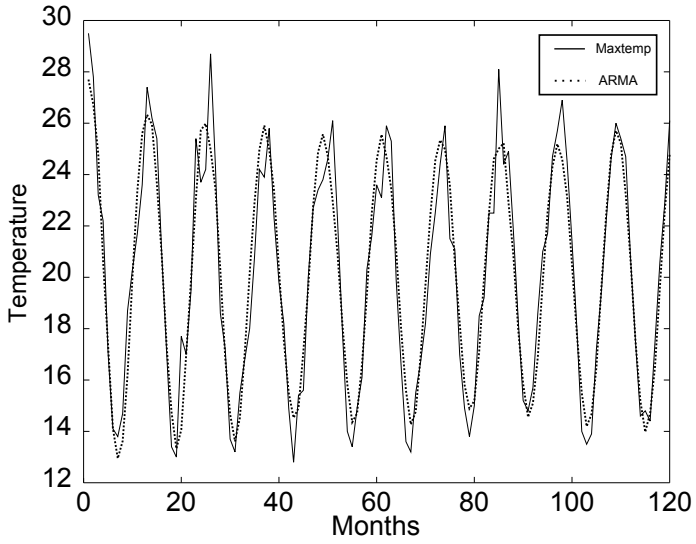
ICDM 2003.

[4] M.Strickert, B.Hammer, Merge SOM for temporal data, Neurocomputing 64:39-72, 2005.

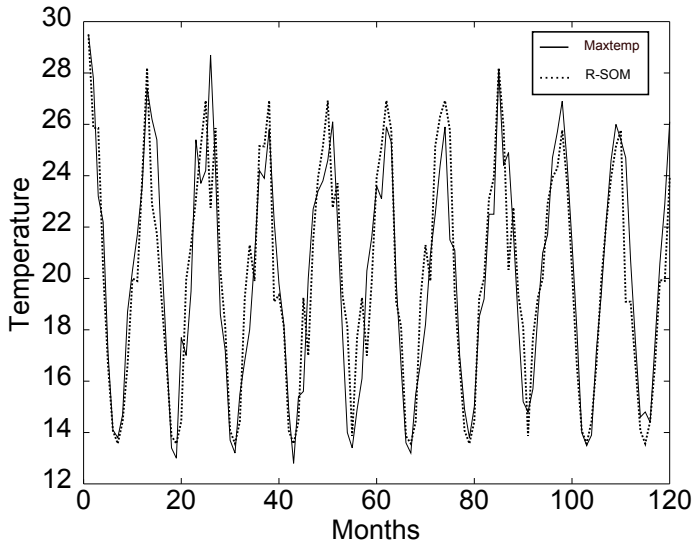
[5] Time Series Data Library

http://www-personal.buseco.monash.edu.au/ hyndman/TSDL/

[6] T. Voegtlin. Recursive self-organizing maps. Neural Networks, 15(8-9):979-991, 2002.



(a) Prediction with ARMA



(b) Prediction with Recursive SOM

Figure 3: One-step prediction on maximum temperature time series

Analysis of Very Large Dissimilarity Datasets

Alexander Hasenfuss[1,2] ***and Barbara Hammer***[1]

1 Introduction

Prototype-based techniques, like k-Means clustering, are widely used in practice. The algorithms combine several striking properties: They are fast, intuitive and very easy to implement, the resulting classes are represented by their means and simple to interpret, and, despite their simplicity, the algorithms are quite powerful. However, the most popular of these algorithms, k-Means, has several drawbacks when applied to typical problems in e.g. computational biology: it optimizes the quantization error by means of a simple batch optimization scheme similar to Expectation Maximization algorithms. It is well known that these methods easily get stuck in local optima of the cost function such that multiple restarts or computationally complex metaheuristics such as genetic algorithms or simulated/deterministic annealing have to be used.

In addition, it depends heavily on the underlying Euclidean metric and yields inappropriate results if the standard Euclidean metric is not suited (e.g. for gene expression data where up- and down-regulations are more important than the absolute values) or not applicable at all.

Several clustering algorithms which partially incorporate these issues and which are based on a cost function similar to the quantization error have been proposed in the context of prototype-based clustering: neighborhood cooperation of neural gas networks [3, 7, 8] is included to avoid local optima; (potentially fuzzy) label information can be incorporated into the clustering by means of an extension of the cost function [8, 17]; the so-called generalized median allows to apply batch clustering to general distance matrices [3, 18]. Recently, Relational Neural Gas was presented [13], a method able to update prototypes continuously in dissimilarity space.

A common challenge today [9], arising especially in computational biology, image processing, and physics, are huge datasets whose pairwise dissimilarities cannot be hold at once within random-access memory during computation, due to the sheer amount of data.

[1]Clausthal University of Technology, Department of Informatics, Clausthal-Zellerfeld, Germany
[2]E-mail: hasenfuss@in.tu-clausthal.de

In recent years, researchers have worked on so-called single pass clustering algorithms which run in a single or few passes over the data and which require only a priorly fixed amount of allocated memory. Popular methods include heuristics such as CURE, STING, and BIRCH [12, 14, 15] and approximations of k-means clustering as proposed in [11].

In this work, we present a technique based on the Relational Neural Gas approach [13], an advanced congener of Neural Gas [7] for dissimilarity data, that is able to handle this situation by a single pass technique based on patches that can be chosen in accordance to the size of the available random-access memory. This results is a linear time and constant memory algorithm for general dissimilarity data which shares the intuitivity and robustness of NG.

2 Relational Neural Gas

Relational data do not necessarily originate from an Euclidean vector space, instead only a pairwise dissimilarity measure d_{ij} is given for the underlying datapoints $v_i, v_j \in V$. The only demands made on dissimilarity measures are non-negativity $d_{ij} \geq 0$, reflexivity $d_{ii} = 0$, and symmetry $d_{ij} = d_{ji}$, so they are not necessarily metric by nature.

One way to deal with relational data is Median clustering [3]. This technique restricts prototype locations to given data points, such that distances are well defined in the cost function of NG. Batch optimization can be directly tranferred to this case. However, median clustering has the inherent drawback that only discrete adaptation steps can be performed which can dramatically reduce the representation quality of the clustering.

Relational Neural Gas (RNG) [13] overcomes the problem of discrete adaptation steps by using convex combinations of Euclidean embedded data points as prototypes. For that purpose, we assume that there exists a set of (in general unknown and presumably high dimensional) Euclidean points V such that $d_{ij} = \|v_i - v_j\|$ for all $v_i, v_j \in V$ holds, i.e. we assume there exists an (unknown) isometric embedding into an Euclidean space. The key observation is based on the fact that, under the assumptions made, the squared distances $\|w_i - v_j\|^2$ between (unknown) embedded data points and optimum prototypes can be expressed merely in terms of known distances d_{ij}. This allows to reformulate the batch optimization schemes in terms of relational data as done in [13].

Note that, if an isometric embedding into Euclidean space exists, this scheme is equivalent to Batch NG and it yields identical results. Otherwise, the consecutive optimization scheme can still be applied.

Relational neural gas shows very robust results in several applications as shown in [13]. Compared to original NG, however, it has the severe drawback that the computation time is $\mathcal{O}(m^2)$, m being the number of data points, and the required space is also quadratic. Thus, this method becomes infeasible for huge data sets. Recently, an intuitive and powerful method has been proposed to extend batch neural gas towards a single pass optimization scheme which can be applied even if the training points do

not fit into the main memory [1]. The key idea is to process data in patches, whereby prototypes serve as a sufficient statistics of the already processed data. Here we transfer this idea to relational clustering.

3 Patch Relational Neural Gas

Assume as before that data are given as a dissimilarity matrix D. During processing of Patch Relational NG, patches of fixed size are cut consecutively from the dissimilarity matrix D, where every patch is a submatrix of D centered around the matrix diagonal.

The idea of the original patch scheme is to add the prototypes from the processing of the former patch P_{i-1} as additional datapoints to the current patch P_i, forming an extended patch P_i^* which includes the previous points in the form of a compressed statistics. The additional datapoints – the former prototypes – are weighted according to the size of their receptive fields, i.e. how many datapoints do they represent in the former patch.

Unlike the situation of original Patch NG [1], where prototypes can simply be converted to datapoints and the inter-patch distances can always be recalculated using the Euclidean metric, the situation becomes more difficult for relational methods.

In Relational NG prototypes are expressed as convex combinations of unknown Euclidean datapoints, only the distances can be calculated. Moreover, the relational prototypes gained from processing of a patch cannot be simply converted to datapoints for the next patch. They are defined only on the datapoints of the former patch. To calculate the necessary distances between these prototypes and the datapoints of the next patch, the distances between former and next patch must be taken into account, as shown in [13]. But that means touching all elements of the upper half of the distance matrix at least once during processing of all patches, what foils the idea of the patch scheme to reduce computation and memory-access costs.

In this contribution, another way is chosen. In between patches not the relational prototypes itselves but representative datapoints obtained from a so called k-approximation are used to extend the next patch. As for standard patch clustering, the points are equipped with multiplicities. On each extended patch a modified Relational NG is applied taking into account the multiplicities.

Patch Relational Neural Gas

Algorithm

Cut the first Patch P_1
Apply Relational NG on $P_1 \longrightarrow$ Relational prototypes W_1
Use k-Approximation on $W_1 \longrightarrow$ Index set N_1
Update Multiplicities m_j according to the receptive fields

Repeat for $t = 2, \ldots, n_p$
  Cut patch P_t
  Construct Extended Patch P_t^* using P_t and index set N_{t-1}
  Apply modified RNG with Multiplicities $\longrightarrow$ Relational prototypes W_t
  Use k-Approximation on $W_t \longrightarrow$ Index set N_t
  Update Multiplicities m_j according to the receptive fields

Return k-approximation of final prototypes N_{n_p}

4 Summary and Outlook

In this contribution, we presented a special computation scheme based on Relational Neural Gas, that allows to process large dissimilarity datasets, that cannot be hold at once in random-access memory, by a single pass technique of fixed sized patches. The patch size can be chosen to match the given memory constraints. The presented patch version reduces the computation and space complexity with a small loss of accuracy.

In future work, the method will be applied to larger real-world datasets in the context of text processing, musical data mining, and computational biology. Furthermore, a supervision concept as reported in [5] can be integrated. The patch scheme also opens a way towards parallelizing the method as demonstrated in [2].

References

[1] N. Alex, B. Hammer, and F. Klawonn, Single pass clustering for large data sets, *Proceedings of 6th International Workshop on Self-Organizing Maps* (WSOM 2007), 2007.

[2] N. Alex and B. Hammer, Parallelizing single patch pass clustering, Proc. ESANN 2008

[3] M. Cottrell, B. Hammer, A. Hasenfuss, and T. Villmann (2006), Batch and median neural gas, *Neural Networks*, 19:762-771.

[4] T. Graepel and K. Obermayer (1999), A stochastic self-organizing map for proximity data, *Neural Computation* **11**:139-155.

[5] B. Hammer, A. Hasenfuss, F.-M. Schleif, and T. Villmann (2006), Supervised batch neural gas, In F. Schwenker, and S. Marinai (Eds.), *ANNPR 2006, Springer Lecture Notes in Artificial Intelligence* 4087:33-45.

[6] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castren (2003), Trustworthiness and metrics in visualizing similarity of gene expression, *BMC Bioinformatics*, **4:48**.

[7] T. Martinetz, S. Berkovich, and K. Schulten (1993). 'Neural gas' network for vector quantization and its application to time series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569.

[8] T. Villmann, B. Hammer, F. Schleif, T. Geweniger, and W. Herrmann (2006), Fuzzy classification by fuzzy labeled neural gas, *Neural Networks*, **19**:772-779.

[9] Q. Yang and X. Wu (2006), 10 Challenging Problems in Data Mining Research, *International Journal of Information Technology & Decision Making* **5(4)**:597-604.

[10] S. Zhong and J. Ghosh (2003), A unified framework for model-based clustering, *Journal of Machine Learning Research* **4**:1001-1037.

[11] S. Guha, N. Mishra, R. Motwani, L. O'Callaghan (2000). Clustering Data Streams. In *IEEE Symposium on Foundations of Computer Science*, 359-366.

[12] S. Guha, R. Rastogi, K. Shim (1998). CURE: an efficient clustering algorithm for large datasets. In *Proceedings of ACM SIGMOD International Conference on Manag ement of Data*, 73-84.

[13] B. Hammer and A. Hasenfuss, Relational Neural Gas. In J. Hertzberg et al., editors, *KI 2007: Advances in Artificial Intelligence*, Lecture Notes in Artifical Intelligence 4667, pages 190-204, Springer, 2007.

[14] W. Wang, J. Yang, R.R. Muntz (1997). STING: a statistical information grid approach to spatial data mining. In *Proceedings of the 23rd VLDB Conference*, 186-195.

[15] T. Zhang, R. Ramakrishnan, M. Livny (1996). BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the 15th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 103-114.

[16] S. Seo and K. Obermayer (2004), Self-organizing maps and clustering methods for matrix data, *Neural Networks* **17:**1211-1230.

[17] T. Villmann, U. Seiffert, F.-M. Schleif, C. Brüß, T. Geweniger and B. Hammer (2006), Fuzzy Labeled Self-Organizing Map with Label-Adjusted Prototypes, In *Proceedings of Conference Artificial Neural Networks in Pattern Recognition (ANNPR) 2006*, F. Schwenker (ed.), Springer, p. 46-56.

[18] T. Kohonen and P. Somervuo (2002), How to make large self-organizing maps for nonvectorial data, *Neural Networks* **15:**945-952.

Topographic Processing of Very Large Text Datasets

Wibke Börger[1,2] ***and Alexander Hasenfuss***[1]

1 Overview

Twenty month from now, the amount of electronic data stored worldwide will be doubled – and the rate is accelerating. As a consequence, nowadays almost every scientific discipline is facing the problem to handle huge data repositories. Here, automatic data mining constitutes an indispensible tool bridging the gap between available data and desired knowledge which would otherwise be inaccessible. Some of the big challenges of real-world data mining have been identified in a panel discussion at the 2007 SIAM International Conference on Data Mining as follows [9]: Mining massive data which go beyond the capacities of standard algorithms, mining streaming data which is generated in a continuous process and which requires immediate feedback, mining heterogeneous data which stem from different sources, applicability of methods and interpretability of the results by researchers outside the data mining community, among others. These facts pose particular requirements towards standard data mining tools concerning their efficiency, flexibility, and interpretability.

Popular large scale applications which combine the aspects of data visualization and clustering are given by the WebSOM which has been used to arrange more than 7 million patents in the visualization space, among other applications [11], and the HSOM which has been used to arrange a large sample of texts from an internet discussion on the flexible hyperbolic space for intuitive visualization [13].

These procedures manage huge document collections by means of growing models and subsampling, respectively. However, they severely rely on the fact that all data are available prior to classification because full information is required for preprocessing of data and for an adequate iterative update of the models. Therefore, the methods cannot be used for huge streaming data sets for which at most a single pass over the data is affordable or which are available only online, respectively. Much effort has been

[1]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[2]E-mail: wibke.boerger@tu-clausthal.de

done to extend clustering methods to streaming data and various extensions of k-means clustering with or without guarantees on the quality of the result and the required recourses [4, 3]. A particularly intuitive and efficient approach has been proposed in [3] for simple k-means: data are subsequently processed in patches whereby the already processed data are represented by means of representative prototypes. This way, a sufficient statistics of data is kept during training and life-long training becomes possible. The approach has recently been extended to more powerful and robust clustering methods such as neural gas [1]. In this contribution, we present the first application of this approach towards data visualization by means of self-organizing maps.

Usually, text data are processed in form of real-valued vectors given by the relevant words in the documents. This procedure requires complex preprocessing of texts and linguistic knowledge, such as stop-word lists and word stems. We will use this compression distance to obtain a general representation of text in this contribution.

The fact that data are represented in terms of a dissimilarity matrix, however, causes two severe problems. On the one hand, powerful clustering methods such as the self-organizing map have been proposed for vectorial data only. We will solve this problem by relying on a recent extension of topographic maps towards general dissimilarity data, as proposed in [8]. On the other hand, clustering and visualization algorithms scale quadratic with the number of data points since the information is represented in a (quadratic) distance matrix.

In this contribution, we solve this problem by transferring patch clustering to the given setting. This way, only constant memory is required and the number of distances used for clustering is only a linear fraction of the quadratic distance matrix, i.e. clustering of quadratic distance matrices in linear time becomes possible. This reduction can be obtained because prototype-based methods allow to substitute already processed data by representative prototypes, i.e. reasonable statistical averaging is automatically included in the method.

Often, the interpretability of unsupervised clustering and visualization is questionable, since the methods extract statistical information of the data without any further guidance by experts in the field. As stated in [10], it is not clear whether extracted information is solely due to noise, or, if not, it is just a trivial statistical effect, or a feature the analyst is simply not interested in. An elegant way to get around this problem has been proposed in [10, 6, 5]: additional information such as a prior labeling is given by the analyst such that visualization methods display only those parts of the data which are relevant to the analyst. The resulting semi-supervised methods offer an intuitive scheme for data clustering and visualization which integrate prior knowledge to the general problem. We show that these methods can be transferred to topographic maps for distance data which are trained in patch mode, such that the interpretability of the models can be guaranteed.

Note that text data usually possess a quite complex underlying topology in correspondence to the complex compositionality of language. Thus a rich topological space should be chosen for visualization. We achieve this goal by means of a visualization in hyperbolic space, as proposed in [13]. This way, we obtain a very flexible topo-

graphic mapping of huge data sets which can deal with arbitrary data structures due to the general compression distance and which can incorporate prior knowledge by means of semi-supervised training.

A sketch of the algorithm used, details can be found in [7]:

Patch Relational Neural SOM

Cut the first Patch P_1 and construct label matrix L_1
Apply Supervised Relational SOM on P_1 incorporating L_1
$\longrightarrow$ Relational prototypes W_1 and learned labels $\bar{L}_1$
Use k-Approximation on $W_1 \longrightarrow$ Index set N_1
Update Multiplicities m_j according to the receptive fields

Repeat for $t = 2, \ldots, n_p$
Cut patch P_t and construct label matrix L_t
Construct Extended Patch P_t^* using P_t and index set N_{t-1}
Construct Extended Patch Labels L_t^* from $\bar{L}_{t-1}$ and L_t
Apply modified SRSOM with Multiplicities on P_t^* incorp. L_t^*
$\longrightarrow$ Relational prototypes W_t and learned labels $\bar{L}_t$
Use k-Approximation on $W_t \longrightarrow$ Index set N_t
Update Multiplicities m_j according to the receptive fields

Return k-approximation of final prototypes N_{n_p} and label statistics $\bar{L}_{n_p}$

2 Large Newsgroup Dataset

As an example for a very large dataset, we gathered 183,546 newsgroup articles from 13 different newsgroups. The text documents were compared by the popular Normalized Compression Distance (NCD) [2], a measure based on approximations of the Kolmogorov Complexity from algorithmic information theory [12]. For our experiments the bzip2 compression method was used.

The full dissimilarity matrix of normalized compression distances for this dataset would occupy approx. 251 GB (!), so it were no option to process it with standard batch methods. Instead, we precalculated NCDs for 183 patches of around 1000 documents each, these dissimilarity matrices were stored to files on hard disk. We then applied the novel Patch Relational SOM with 3-approximation and a hyperbolic grid of 85 neurons. That way, the computation time was around 18h instead of an extrapolated half a year! Also the computation required only a constant space of some megabytes and could be performed on a common workstation.

Most time consuming part of the calculation was the construction of the extended patch, here we had to determine the normalized compression distances between neurons and datapoints on the fly. Due to the size of the problem it is not possible to calculate and store those distances in advance.

The outcome is a mapping into 2-dimensional hyperbolic space, that can be projected to the Euclidean plane for visualization and data inspection (see fig. 4).

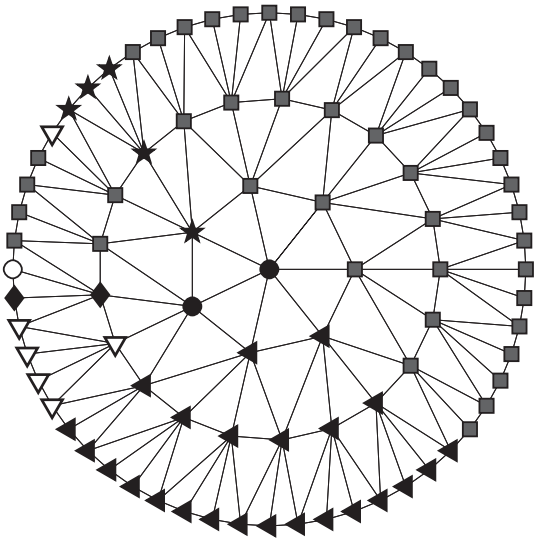


Figure 4: Visualization of 183,546 newsgroup articles Hyperbolic Patch RSOM

3 Conclusions

In this work, we presented a (semi-)supervised variant of Patch Relational Neural Gas and introduced the novel (semi-)supervised Patch Relational SOM. Both methods are suited for processing of very large dissimilarity datasets, especially arising in the field of text mining. Due to the fact, that new incoming patches can arbitrarily be processed later on, the methods are especially suited for long-term learning.

Further work shall include the development of a toolbox providing more sophisticated visualization methods – as available for standard SOM – also for the Patch Relational SOM.

Also under development are applications of the patch relational methods on more real world text data sources, like e.g. web directories or musical notation databases. These methods are designed to be part of modern information systems, like visual search engine, especially making use of the long-term learning capabilities and the ability to handle large amounts of data.

References

[1] N. Alex, B. Hammer, and F. Klawonn, Single pass clustering for large data sets, *WSOM*, 2007.

[2] R. Cilibrasi and M.B. Vitáni (2005) Clustering by compression, *IEEE Transactions on Information Theory* 51(4):1523-1545.

[3] F.Farnstrom, J.Lewis, C.Elkan (2000), Scalability for clustering algorithms revisited, *SIGKDD Explorations* 2(1):51-57.

[4] S. Guha, N. Mishra, R. Motwani, L. O'Callaghan (2000). Clustering Data Streams. In *IEEE Symposium on Foundations of Computer Science*, 359-366.

[5] B. Hammer, A. Hasenfuss, F.-M. Schleif, and T. Villmann (2006), Supervised batch neural gas, In *Proceedings of Conference Artificial Neural Networks in Pattern Recognition (ANNPR)*, F. Schwenker (ed.), Springer, pages 33-45.

[6] B. Hammer, A. Hasenfuss, F.-M. Schleif, and T. Villmann (2006), Supervised median neural gas, In Dagli, C., Buczak, A., Enke, D., Embrechts, A., and Ersoy, O. (Eds.), *Intelligent Engineering Systems Through Artificial Neural Networks* 16, Smart Engineering System Design, pp.623-633, ASME Press.

[7] A. Hasenfuss, W. Boerger, B. Hammer (2008), Topographic Processing of Very Large Text Datasets, submitted to ANNIE 2008.

[8] A. Hasenfuss and B. Hammer (2007), Relational Topographic Maps, Proc. IDA 2007, Lecture Notes in Computer Science 4723, 93–105.

[9] H. Hirsch (2008), Panel Summary, Data Mining Research: Current Status and Future Opportunities, *Statistical Abalysis and Data Mining*, DOI: 10.1002/sam.10003.

[10] Kaski, S., Sinkkonen, J., and Peltonen, J. (2001). Learning metrics for self-organizing maps. In Proceedings of the International Joint Conference on Neural Networks, volume 2, pages 914–919.

[11] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela (2000), Self Organization of a Massive Document Collection. IEEE Transactions on Neural Networks, 11(3):574-585.

[12] M. Li and P. Vitányi (1997), *An Introduction to Kolmogorov Complexity and Its Applications*, Springer.

[13] J. Ontrup and H. Ritter (2001), Hyperbolic self-organizing maps for semantic navigation. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14 (NIPS), volume 14, pages 1417–1424, Cambridge, MA, 2001. MIT Press.

TU Clausthal
ICOLE 2008, LESSACH, AUSTRIA

A Novel Dissimilarity Measure for the Topographic Mapping of Symbolic Musical Data

Bassam Mokbel[1,2]***, Alexander Hasenfuss***[1]

1 Introduction

The ever increasing amount of music collections available in online stores or public databases has created a need for user-friendly and powerful interactive tools which allow an intuitive browsing and searching of musical pieces. Ongoing research in the field of music information retrieval thus includes the adaptation of many standard data mining and retrieval tools to the music domain. In this regard the topographic mapping and visualization of large music compilations combines several important features: data and class structures are arranged in such a way that an inspection of the full dataset as well as an intuitive motion through partial views of the database become possible.

Generally, there are two basic ways to automatically construct the topographic arrangement for a mapping: The first is to use a set of n features to position each subject in an n-dimensional space. Then, since n is usually much larger than 3, the higher-dimensional map has to be projected to 2 or 3 dimensions for visualization with e.g. linear mapping methods like MDS or PCA. This causes a certain amount of information loss and distortion. Also, to put the subjects to certain coordinates it is necessary to use global features that are valid on the entire dataset. The second method uses pairwise dissimilarities between all subjects for positioning. This avoids the aforementioned disadvantages like the additional effort for dimensionality reduction, which is no longer necessary. Instead, the positions can be calculated directly out of the distances. However, if the given pairwise dissimilarities are not isometrically embeddable into Euclidean space, a projection has to be distorted as well. Also, due to its restriction to Euclidean data, the classical self-organizing map (SOM) as introduced by Kohonen in [6] cannot be used here. Further, it seems doubtful in principle that a standard rect-

[1]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[2]E-mail: bassam.mokbel@gmail.com

angular SOM constitutes the right space to display complex musical collections which usually show a much richer neighborhood structure than a Euclidean one.

We therefore apply two extensions of the classical SOM to the music domain which overcome these problems: The Relational SOM, a model that has recently been proposed in [4], can directly deal with arbitrary similarity data instead of only simple Euclidean ones. On the other hand, we use a non-Euclidean hyperbolic SOM with a hyperbolic grid as visualization space that is capable of representing a much richer topology and neighborhood structure characteristic for musical data collections, as described in [10].

For music, there are many different features that can be extracted by algorithmic methods, either from acoustic or symbolic data of a musical piece. To compute a dissimilarity between pieces based on their tonal and rhythmic progression, it is possible to use the temporal progression of features like rhythmic patterns, note or chord sequences to measure dissimilarity. This can be achieved with a suitable method to measure string dissimilarity like the edit distance or the powerful and universal compression distance. Especially the latter has been used in this way on symbolic representations of musical data with promising results in recent years, like e.g. in [1, 3, 8].

Due to the nature of acoustic signals, it usually requires much more effort to extract high-level features from acoustic audio data than from symbolic music representations like MIDI or MusicXML. But recent progress in developing efficient and reliable automated extraction methods that are able to gain musical notation directly from complex acoustic material, as presented in [5, 13, 9], open the way towards mapping techniques which directly rely on a symbolic description of musical data.

In the following section, we introduce a new way to convert such symbolic representations into strings via a priorly constructed precedence structure. We implemented our method in Matlab and used MIDI files as symbolic input data. We processed selected subsets as well as the entire archive of classical pieces from the *Kunst der Fuge*[1] MIDI collection consisting of about 12500 files. The generated dissimilarities were mapped by a Supervised Relational Hyperbolic SOM in a hyperbolic grid and also with a non-metric multidimensional scaling with Kruskal's normalized stress1 criterion. The experiments show most of the data arranged in meaningful clusters with a reasonable separation of composers and eras.

2 Invariant Dissimilarities for Symbolic Musical Data

To measure the dissimilarity between the tonal or rhythmic progression of two musical pieces, our method compares string representations derived from them. We therefore developed an algorithm that converts the symbolic note sequences in a MIDI file into a string, following a priorly constructed precedence structure. Our algorithmic approach is based on the assumption that a human's subjective perception of musical identity usually works very context-driven. Thereby we suppose that most listeners will consider a

[1] `http://www.kunstderfuge.com`

melody to a certain extent similar to a copy of it, if it has been changed in the following ways:

- It is shifted in its overall pitch, i.e. it has been transposed to another fundamental note.
- It is scaled in its overall tempo, i.e. all note lengths and pauses have been contracted or elongated by a constant factor.

Thus, *transposition-invariance*[2] and *time-scale-invariance* is beneficial for an automatic comparison based on symbolic representations. Under this precondition, those parts of two musical pieces that are in the aforementioned sense similar or equal, would yield equal parts in their string representations calculated by our method. The information that is not encoded in these strings is the magnitude by which the pitch was shifted or the tempo scaled. As the described human assessment of similarity would probably decrease along with a raise in magnitude of such changes, it might be more truthfully described by distinguishing degrees of similarity. Although this is not part of our encoding scheme at the moment, it is easily imaginable to incorporate such information into the measure in the future.

Some related methods can be found in literature that partially provide for the emphasized invariances, like e.g. [11, 12, 3, 8]. In [3] transposition-invariance was achieved with a global pitch normalization throughout the entire piece, making the encoding very sensitive to the automatic choice of the global point of reference. In [8] every note's pitch was encoded as the difference to the pitch of its directly preceding note. In addition to independence of the overall pitch, this method yields local separation: parts in the strings are equal for parts of two songs that, aside from transposition, have equal note sequences, even if the rests of these songs are completely different. Using common string dissimilarity measures on those two representations would therefore reflect the partly equality in its output value. In addition, one could store note lengths and pauses analogously to gain time-scale-invariance. Still, in these strings you will find only very little equality in the case of two songs playing the same melody (like a riff or a theme), only with dissimilar orchestration and arrangement surrounding it. Then the output of a string dissimilarity measure would in our opinion not represent most listeners' assessments.

Our goal for the generated string representation was therefore a decomposition of the tonal and rhythmic progression of a song, that has the benefit of local points of reference, but, on top of that, represents riffs and themes more independently of the surrounding melodic context. Our strategy is to automatically define precedence relationships between notes throughout the entire piece:

The functions $start(n)$, $pitch(n)$ and $length(n)$ return the start time, pitch and length of a note n respectively. For every note n of all played notes N, the algorithm picks one designated predecessor. For the current note $cn \in N$, the function $pred(cn)$

[2] also referred to as *pitch-invariance*

returns one of its time-wise preceding notes on the same MIDI channel as the predecessor note (an element of $P(cn)$).

We define

$$pred(cn) = \underset{p \in P(cn)}{\operatorname{argmin}} \left(k \cdot |pitch(p) - pitch(cn)| + r \cdot \frac{start(cn) - start(p)}{length(cn)} \right)$$

with $P(n) = \{ x \in N : start(x) < start(n) \}$ and $p \in P(cn)$, $cn \in N$.

The predecessor is thereby chosen to be the closest prior note in terms of the difference in start time *and* in pitch. Since the time-wise distance is calculated relatively to the current note's length, the field of search is stretched or shrinked in the time-dimension depending on its length. As an alternative strategy one could also consider stop times of prior notes instead. The global parameters k and r control the overall search strategy.

The resulting tree structures of precedences, as seen in an example in figure 5, are then utilized to store the change in pitch and length as well as the difference in note start times at every edge along every path of every tree. For a note n the named changes are calculated and stored in relation to its predecessor $pr = pred(n)$ as follows:

$$relpitch(n) = pitch(n) - pitch(pr),$$
$$reltiming(n) = \frac{start(n) - start(pr)}{length(pr)}, \qquad rellength(n) = \frac{length(pr)}{length(n)}.$$

Next to $rellength(n)$, the value of $reltiming(n)$ adds some more information about the rhythmic expression in the string representation, as they conjointly also encode the existance and length of pauses in between the notes.

Considering the precedence structure, a small, local change in the musical progression will subsequently cause it to alter locally, resulting yet only in a local symbolic change of the string representation. To explain the benefit of this behavior on a practical example, imagine a bass line which is due to its lower pitch isolated in its tonal progression. Its melody will be independently represented in the string, invariant to changes of higher lead melodies played at the same time on the same MIDI channel. If those lead lines are sufficiently pitch-wise separated from each other as well, they will also have independent representations within the string.

To sum it up, our encoding method is fully transposition-invariant and time-scale-invariant on an entire song but also shows highly invariant behavior upon changes on every subset of notes in the song. Furthermore, even a certain amount of invariance to variations of melodies is achieved.

To calculate the dissimilarity of the string representations, we used the popular Normalized Compression Distance (NCD) (see [2]), a measure based on approximations of the Kolmogorov Complexity from algorithmic information theory described in [7].

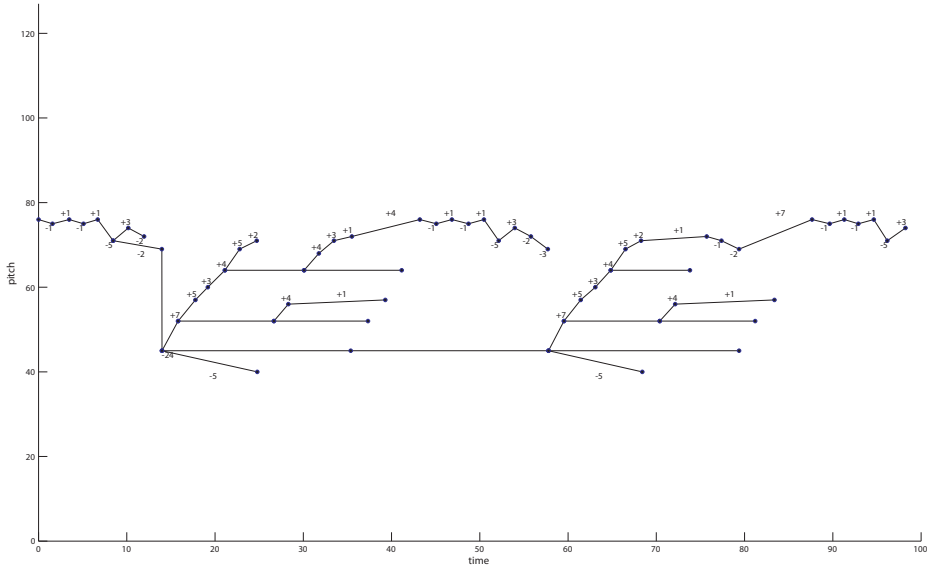


Figure 5: The precedence tree structure for the first 60 notes of Beethoven's "Für Elise". The edges are marked with all nonzero pitch changes of notes relative to their predecessors.

For our experiments the bzip2 compression method was used. Since bzip2 works byte-oriented as most of the common compression methods, the size of a reasonable set of symbols is restricted to 2^8. Therefore, we utilize the integer values in $[1..255]$ to code every possible relative state change of a note compared to its predecessor. For every musical piece, we automatically build two strings, one that holds the pitch changes $relpitch(n)$ for every note n and another one for the rhythmic progression. The latter is compiled from $rellength(n)$ and $reltiming(n)$, resulting in a string which is twice as long as the one representing the pitches. The dissimilarity of two songs is then calculated as a weighted mean of the NCD of the pitch strings and the NCD of the rhythmic strings.

References

[1] Z. Cataltepe, Y. Yaslan, and A. Sonmez. Music genre classification using midi and audio features. *EURASIP Journal on Advances in Signal Processing*, page Article ID 36409, 2007.

[2] R. Cilibrasi and P. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, April 2005.

[3] R. Cilibrasi, P. Vitányi, and R. de Wolf. Automatic clustering of music based on string compression. *Computer Music Journal*, 28(4):49–67, 2004.

[4] B. Hammer and A. Hasenfuss. Relational neural gas. In J. Hertzberg, M. Beetz, and R. Englert, editors, *KI 2007: Advances in Artificial Intelligence*, volume 4667 of *Lecture Notes in Artificial Intelligence*, pages 190–204, Berlin, 2007. Springer.

[5] A. Klapuri and M. Davy, editors. Springer, New York, 2006.

[6] T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

[7] M. Li and P. Vitányi. *The Dissimilarity Representation for Pattern Recognition*. Springer, 1997.

[8] A. Londei, V. Loreto, and M. O. Belardinelli. Musical style and authorship categorization by informative compressors. In *Proceedings of the 5th Triennial ESCOM Conference*, September 2003.

[9] B. Pardo and W. P. Birmingham. Algorithms for chordal analysis. *Computer Music Journal*, 26(2):27–49, 2002.

[10] H. Ritter. Self-organizing maps in non-euclidean spaces. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 97–108. 1999.

[11] A. Ruppin and H. Yeshurun. Midi music genre classification by invariant features. In *Proceedings of the International Conference of Music Information Retrieval*, 2006.

[12] E. Ukkonen, K. Lemström, and V. Maekinen. Geometric algorithms for transposition invariant content-based music retrieval. In *Proceedings of the International Conference of Music Information Retrieval*, 2003.

[13] J. Woodruff and B. Pardo. Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings. *EURASIP Journal on Advances in Signal Processing*, page Article ID 86369, 2007.

TU Clausthal
ICOLE 2008, LESSACH, AUSTRIA

Core Knowledge for a Humanoid Robot Based on Findings in Infant Research

Jan Biel[1,2]

1 Introduction

Converging evidence from the field of cognitive science promotes the assumption that certain core abilities might be available to the human infant from birth. One of these is the ability to discriminate between inanimate objects and agents. Furthermore, studies with infants show that the ability to infer the goal of an agent is available to the human child within its first year of life. These findings provide an interesting interface with the field of robotics. Endowing a robot with basic core abilities can help to form the basis of a higher cognitive apparatus able to infer an agent's more complex goals. Such a mechanism can help the robot to better navigate the complex world of humans. With only little research in that area, the graduation thesis [Bie08], developed at the Honda Research Insititute Europe GmbH, attempts to take a first step to pave the way toward further investigations.

2 Results from Cognitive Science

During the last decades, cognitive science has discovered that human infants have more sophisticated perceptional skills than previously assumed. By employing a *looking time* paradigm typical for measuring preverbal infants' abilities, some interesting results were derived. Many of these studies deal with the perception of geometrical shapes that do not display any recognizable features like e.g. the human hand.

Six-month old infants are able to perceive the collision of two objects as causal, but not if there is a spatial or temporal gap involved [LK87]. This result leads to a model according to which, in infant perception, inanimate objects only interact through contact, while animate objects (or agents), can interact without contact [Les93]. Furthermore, by

[1]Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[2]E-mail: jan.biel@arcor.de

the age of twelve months, infants are able to interpret an action as goal-directed, identifying goals of an actor and constraints under which these goals are reached [Ger03].

3 Modeling Infants' Abilities

For transferring the discrimination between inanimate objects and agents, as well as the perception of an agent's simple goals to a robot, adequate models were required. To test the developed techniques, scenes needed to be generated that display situations similar to the experimental setups from infant studies.

To that end, a simulation environment was developed using the Open Dynamics Engine (ODE)[1], which natively handles the collisions necessary in this context. To generate scenes in which agents follow simple goals, a model based on potential fields was developed in which each object in a scene acts as an attractor or a repellor to an agent, effectively defining the agent's movement.

The model for discriminating between agents and non-agents employs a collision and velocity change detector, following the idea that objects interact through collisions, while agents can interact without collisions. Since the visual data is usually subject to noise, a collision rating based on the distance between two objects is used. The velocity change is measured by using a Kalman Filter with a constant velocity model. The square difference between the prior and posterior velocities measures the velocity change. Both detectors map their result to the interval [0..1] to make them comparable.

In order to measure goals of an agent, a model similar to the one in the scene generation is considered. Again, each object acts as the source of a potential field determining the motion of the agent. By measuring the agent's movement and the objects' positions at only one time step, the distribution of potential strengths for each object can be computed. This is achieved by taking into account the model equation in dependence of the unknown potential strengths. By minimizing the error between the hypothesized and measured movement of the agent, the most plausible potential strengths can be derived. This minimization problem leads to an equation system which can be solved by Gaussian elimination.

Singularities in the equation system can be detected by using the condition number of the coefficient matrix and excluded from consideration.

Experiments showed that when dealing with more than three objects, the linear equation system becomes singular in every case, which makes getting a solution impossible for these types of scenes. Research from cognitive science has shown that infants are also unable to represent more than three objects, while adults can follow up to four objects moving on a continuous path [SK07]. To see how the model holds up to these problematic cases, two extensions have been investigated.

The first extension takes into account more than one measurement in contrast to the basic model which only considered one measurement per time step. This modification extends the linear equation system to an overdetermined equation system which can be

[1] `http://www.ode.org`

solved by applying the least squares method. By calculating the potential strengths from two measurements, the singularities in the three-object case can be largely avoided, while even a previously not evaluable case with five objects can be fairly well interpreted.

The second extension simplifies the problem by assuming that only one object significantly influences the agent's movement, while all remaining objects act as obstacles with the same repelling potential. This modification results in a simplified overdetermined equation system which is again solvable by least squares. Like the first extension, this modification of the original model allows to interpret scenes with more than three objects, with decent results that still leave room for improvement.

References

[Bie08] Jan Biel. Core knowledge for a humanoid robot based on findings in human infant research. Diplomarbeit, Clausthal University of Technology, Department of Informatics, 2008.

[Ger03] György Gergely. What should a robot learn from an infant? mechanisms of action interpretation and observational learning in infancy. *Connection Science*, 15(4):191–209, 2003.

[Les93] Alan M. Leslie. A theory of agency. Technical report, Rutgers University Center for Cognitive Science, 11 1993.

[LK87] Alan M. Leslie and Stephanie Keeble. Do six-month-old infants perceive causality? *Cognition*, 25(3):265–288, 1987.

[SK07] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

Reading a DNA Sequence - From Sequencing to Assembling

***Aleksandra Swiercz*[1,2,3], *Jacek Blazewicz*[1,2], *Marek Figlerowicz*[2], *Piotr Gawron*[1] *and Marta Kasprzak*[1,2]**

Acknowledgements: The work has been partially supported by the Polish Ministry of Science and Higher Education grant N N519 314635.

1 Introduction

Analysis of DNA sequences has become an essential issue over the past few years, and has been developed by many research projects in molecular biology. Till now many genomes were already sequenced, but still many are unknown. One cannot read a sequence of nucleotides directly. Thus, the process of recognizing the genetic information is divided into three steps: sequencing, assembling and mapping. In the first step, sequencing, one can obtain the fragments of length up to a few hundreds of nucleotides. Different biochemical approaches were proposed, the most known are Sanger method [5, 6] and sequencing by hybridization (SBH) [7, 2]. Yet all of the methods produce errors. Recently, new approaches were developed which generate quite reliable fragments in shorter time and at a lower cost. First approach is based on the pyrosequencing protocol and was proposed by 454 Life Sciences Corporation. Another platform introduced by Illumina – the Solexa technology – is based on massively parallel sequencing of millions of fragments at the same time. The fragments are much shorter than obtained by the other approaches, and the method is mostly used for resequencing.

The next step for reading a DNA sequence is the assembling. Its aim is to combine fragments obtained during the sequencing phase into a longer sequence of length up to millions of nucleotides. The problem is known from its high complexity, due to huge amount of erroneous data.

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]Institute of Bioorganic Chemistry, Noskowskiego 12/14, 61-704 Poznan, Poland
[3]E-mail: aswiercz@cs.put.poznan.pl

In the last step the genome is completed either by mapping DNA fragments on the proper place on the chromosome or by experts' finishing. For smaller genomes the last step can be omitted.

The aim was to compare existing approaches to DNA sequencing and assembling and to propose a new method for assembling short reads which uses the concept of the modified graph developed in the context of sequencing by hybridization.

2 DNA sequencing

Pyrosequencing is a novel method of DNA sequencing developed by Mostafa Ronaghi ([1]). This method allows for a sequencing of a single strand of DNA by synthesizing a complementary strand along it. Each time after a nucleotide A, C, G, or T is incorporated into a newly created chain, many cascade enzyme reactions are triggered what at last ends in light emission. The number of nucleotides of one type, which joins the complementary sequence in one step, depends on the number of the consecutive nucleotides of the same type in the sequence. The light signal is proportional to the number of nucleotides incorporated and is detected on the camera device. After each cycle the rest of not joined nucleotides are washed up and the process is continued with another nucleotide. It is possible to analyze a large number of samples in the same time. The pyrosequencing process is fully automated and with relatively low cost comparing to other sequencing methods. However, this method has serious limitations with short reads, which makes the process of genome assembling much more complicated.

Recently 454 Life Sciences company has applied the pyrosequencing approach for assembling short genomes [4]. In this approach the step of sequencing is done automatically. The 454 sequencer is able to resolve hundreds of thousands of nucleotides per one run.

3 Assembling

As the output of the sequencing phase one gets many DNA fragments of different lengths which come from both strands of a DNA helix. In general, the sequences may contain errors: insertion (additional nucleotide which do not appear in the original sequence), deletion (lack of a nucleotide in the fragment), and substitution (replacement of a proper nucleotide by the other one). The assembly problem becomes difficult because imperfect matches have to be allowed while aligning the fragments together.

If there are no errors in the input set of fragments then we get the shortest common superstring problem, which is known to be NP-hard [3]. In case of errors in the fragments the problem is even more difficult to solve.

The DNA assembly problem in case of errors can be formulated as follows: (optimization version)

Instance: Multiset S of fragments over the alphabet A,C,G,T, which can come from both strands of DNA helix. All the fragments reverse and complementary to the original

are added to the set S.

Goal: The sequence of the highest value of probability, containing from each pair of the fragments in the set (original fragment or the reverse complementary one) only the one fragment as a subsequence (small errors in the alignment of the fragments are allowed).

The novel algorithm was proposed for assembly of the short reads coming from 454 sequencer. The method is the heuristics based on a graph model. The new features which were developed are the compression of the input sequences, very fast multiple alignment heuristics. The usefulness of the algorithm has been proved in tests on raw data generated during sequencing of the genome of bacteria Prochlorococcus marinus.

References

[1] M. Ronaghi and S. Karamohamed and B. Pettersson and M. Uhlen and P. Nyren: Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem., 1996, 242, 1, p. 84-89

[2] W. Bains and G.C. Smith: A novel method for nucleic acid sequence determination. J. Theoretical Biology, 1988, 135, p. 303-307, BS88

[3] J. Gallant and D. Maier and J.A. Storer: On finding minimal length superstrings. J. Comput. System Sci., 1980, 20, p. 50-58

[4] M. Margulies and M. Egholm and W.E. Altman and S. Attiya and others: Genome sequencing in microfabricated high density picolitre reactors. Nature, 2005, 437, p. 376-380

[5] A.M. Maxam and W. Gilbert: A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA, 1977, 74, p. 560-564

[6] F. Sanger and S. Nickelen and A.R. Coulson: DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA, 1977, 74, p. 560-564

[7] E.M. Southern, United Kingdom Patent Application GB8810400, 1988

TU Clausthal
ICOLE 2008, LESSACH, AUSTRIA

A New Approach to DNA Assembly

Piotr Gawron*[1,3]*, Jacek Blazewicz*[1,2]*, Marek Figlerowicz*[2]*,

***Marta Kasprzak*[1,2] *and Aleksandra Swiercz*[1,2]**

Acknowledgements: The work has been partially supported by the Polish Ministry of Science and Higher Education grant N N519 314635.

Progress in bioengineering brought a new approaches to DNA sequencing, which aim is to give highly reliable output of low cost and in short time. These new approaches, like for example 454 sequencing, provide much more data in shorter time. Because of the sequences reliability this method is much better than others for assembly purposes. However, produced sequences are much shorter and there are many more of them, which indicate that the problem is harder. Presented algorithm was created to process data from 454 sequencing method.

The algorithm is a heuristic. Solution bases on graph model. The problem was divided into few subproblems: creating a graph, modifying the graph and finding a path within it, and extracting a genome sequence from the path. Classical, exact solutions to all of these subproblems are very time-consuming. In presented algorithm heuristic methods were used to solve them, which compute faster and still return high quality solutions. During the first phase (creating the graph) the input sequences become vertices, and the connections between vertices (edges) are created on the base of the high heuristic note, which indicates probability of the alignment between sequences. Finding a path containing all vertices (or few few disjoint paths containing all vertices) in graph is a subproblem of Traveling Salesman Problem, which in general is NP-hard. As a solution to this subproblem a greedy algorithm was used. Extracting final sequence from the path can be transformed to multialignment problem, which is NP-hard too. However there are few constraints which makes this problem easier: input sequences are overlapping each other very tightly and the order of sequences is known, thus a new heuristic method was proposed which bases on this knowledge and computes results with very good quality.

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]Institute of Bioorganic Chemistry, Noskowskiego 12/14, 61-704 Poznan, Poland
[3]E-mail: `piotr.gawron@cs.put.poznan.pl`

Usefulness of the algorithm has been proven in tests on raw data generated during sequencing of the whole 1.84 Mbp genome of bacteria Prochlorococcus marinus.

TU Clausthal
ICOLE 2008, LESSACH, AUSTRIA

Constructing Oligonucleotide Libraries Based on Graph Theory Models

Adam Kozak[1,3]***, Tomasz Glowacki***[1]***, Piotr Formanowicz***[1,2]

Acknowledgements: The work has been partially supported by the Polish Ministry of Science and Higher Education grant N N519 314635.

1 Introduction

DNA libraries are usually built of chains that represent real genetic information. These libraries are obtained from biochemical experiments and provide some information for researchers (e.g. cDNA library consists of DNA translated from mRNA chains and allow to determine gene expression). There are some experiments that do not require natural genetic material.

In 1994 Adleman showed that DNA can be used also for computing purposes and presented a way to find Hamiltonian path in a graph [1]. Vertices in such a graph were encoded as randomly generated DNA chains of length l. Arcs were encoded as DNA chains that consist of last $\frac{l}{2}$ nucleotides of preceding vertex and first $\frac{l}{2}$ nucleotides of succeeding vertex. Hybridization between vertices and arcs represented path in a given graph. DNA chains that represented vertices were generated randomly, so they could also hybridize with each other.

The goal of this work is to present an algorithmic method of constructing DNA libraries that would be able to encode such vertices and minimise the probability of hybridization between library elements.

2 Problem description

A library should contain n DNA chains of length k. These chains can be represented in concatenated form as one long chain of length $p = nk$. Library elements have minimal

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]Institute of Bioorganic Chemistry, Noskowskiego 12/14, 61-704 Poznan, Poland
[3]E-mail: adam.kozak@cs.put.poznan.pl

tendency to hybridize with each other, so concatenated chain of length p is called *an internally anticomlementary chain*. DNA alphabet has four letters, but the problem can be formulated for any alphabet of even length (length must be an even number, because only then it is possible to define relation of complementarity).

Problem 2.1. *CIAC – Construction of Internally Anticomplementary Chain*
The following data is given:

1. Alphabet $\Sigma = \{0, 1, \ldots, \alpha - 1\}, \quad |\Sigma| = \alpha, \quad 2|\alpha,$
2. Symmetric binary relation of complementarity $l_1 \sim l_2$ defined on set Σ^2 such, that: $\forall l_k \in \Sigma : \quad |\{l \in \Sigma : l_k \sim l\}| = 1$ and $l_1 \sim l_2 \Rightarrow l_1 \neq l_2$
3. Length of chain p.
4. Symmetric binary relation of anticomplementarity R_j defined on set $\left(\Sigma^j\right) \times \left(\Sigma^j\right)$ such, that:

$$a = \left(l_1^a, \ldots, l_j^a\right),\ b = \left(l_1^b, \ldots, l_j^b\right) \in \Sigma^j$$
$$R_j(a, b) \Leftrightarrow \exists i \in \{1, \ldots, j\} : \neg \left(l_i^a \sim l_i^b\right)$$

Algorithm should build result chain P of length $|P| \geq p$ over alphabet Σ that satisfies the following condition:

$$\forall a, b \subset P, |a| = |b| = j :\ R_j\,(a, b) \tag{1}$$

and each letter of alphabet Σ has equal probability to occur in P.

Algorithm for constructing the described library is based on labeled graphs (Fig. 6). These graphs are used for example in computational biology for assembling data obtained in DNA sequencing by hybridization [4]. Labeled graphs are also adjoints [3] and it is possible to search for Hamiltonian cycle in such graphs in polynomial time [5, 2].

Definition 2.1. ([2]) Let $k > 1$ and $\alpha > 0$ be two integers. Then 1-graph $H(V, A)$ can be (α, k)-labeled if it is possible to assign a label $(l_1(x), l_2(x), ..., l_k(x))$ to each vertex x of H that:

1. $l_i(x) \in \{0, \ldots, \alpha - 1\}$ for all $v \in V$
2. Each label is unique, so $\forall x \neq y : (l_1(x), \ldots, l_k(x)) \neq (l_1(y), \ldots, l_k(y))$,
3. There is an arc between vertices x and y if and only if $k - 1$ last letters of label of vertex x are equal to first $k - 1$ letters of label of vertex y:
   $(x, y) \in E \Leftrightarrow (l_2(x), \ldots, l_k(x)) = (l_1(y), \ldots, l_{k-1}(y))$

Algorithm that solves problem 2.1 has following steps:

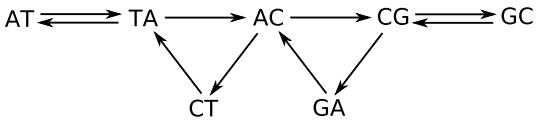


Figure 6: (4,2)-labeled graph

1. Compute length k of vertex label that allow to build chain of length p.
2. Build (α, k)-labeled graph $G(V, A)$ that contains all possible labels of length k (α^k vertices and α^{k+1} arcs).
3. For every pair a_1, a_2 of arcs in $G(V, A)$ such that

   $$a_1 = (v_1, v_2),\ a_2 = (v_3, v_4) \in A\ \wedge\ \neg R_k(v_1, v_3)\ \wedge\ \neg R_k(v_2, v_3)$$

   remove a_1 or a_2. Removing arcs should keep Eulerian cycle in obtained graph $G'(V, A')$.
4. Find Eulerian cycle $E = (v_1, v_2, v_3, \ldots, v_n)$ in graph $G'(V, A')$ ($|A'| = n = \frac{|A|}{2} = \frac{\alpha^{k+1}}{2}$).
5. Build resulting chain P using Eulerian cycle $E = (v_1, v_2, v_3, \ldots, v_n)$:
   $P =$ **empty chain**
   for i=1 to n: $P \leftarrow P + l_k\,(v_i)$

Every pair s_1, s_2 of subchains of P such that $|s_1| = |s_2| = k{+}1$ satisfy condition (1), because they represent arcs in graph $G'(V, A')$. Constructing solution from Eulerian cycle imposes following condition for length of vertex label:

$$|P| = n = \frac{\alpha^{k+1}}{2}\ \Rightarrow\ \alpha^{k+1} = 2\,|P|\ \Rightarrow\ k = \log_\alpha 2\,|P| - 1 = \lceil \log_\alpha 2\,|P| \rceil - 1 \quad (2)$$

Parameter k obtained from equation (2) is a result of the first step of the algorithm – it represents the length of vertex label, so it has to be an integer value. Given the length of chain p does not have to implicate integer value of a logarithm, so it has to be rounded.

3 Conclusions

This work presents a method of constructing DNA libraries that consist of elements that have minimal tendency to hybridize with each other. Constructed libraries can be applied in DNA computing to encode problem instances. Each element of the library is a part of chain constructed by the algorithm. This chain can be used also for other purposes, because it has minimal tendency to hybridise with its own parts.

References

[1] Adleman, L.M.: Molecular computation of solutions to combinatorial problems. Science, 266, 1994, p. 1021–1024.

[2] J. Blazewicz, A. Hertz, D. Kobler and D. de Werra: On some properties of DNA graphs. Discrete Applied Mathematics, 98, 1999, p. 1–19.

[3] C. Berge: Graphes. Dunod, Paris, 1970.

[4] M. Kasprzak: On the link between DNA sequencing and graph theory. Computational Methods in Science and Technology, 10, 2004, p. 39–47.

[5] Ch. H. Papadimitriou: Computational complexity. Addison Wesley, December 10, 1993.

TU Clausthal
ICOLE 2008, LESSACH, AUSTRIA

Comparison of RNA Structures – Concepts and Measures

***Tomasz Zok*[1,3], *Mariusz Popenda*[2], *Marta Szachniuk*[1,2]**

Acknowledgements: The work has been partially supported by the Polish Ministry of Science and Higher Education grant N N519 314635.

1 Introduction

Since the first analytical study of molecular structures, the ability to compare their conformations has been highly important. Molecules sharing structural features may demonstrate similar functional characteristics, thus, investigating similarities within the set of given conformations has been always a scientifically promising procedure. Therefore, it is crucial to provide tools for measuring distances between the structures.

Root mean square deviation (Kenney and Keeping, 1962; Hoehn and Niven, 1985; Kavraki, 2007) is one of the most common measures used to compare tertiary structures of biomolecules. It is based on algebraic representation of the structure, which determines three Cartesian coordinates for each atom of the molecule. Since RMSD strongly depends on reciprocal orientation of structures and requires finding their optimal superposition to minimize the distance, it provided us with the idea of applying the other distance measure. Computing mean of circular quantities is possible due to trigonometric representation of the structure and does not require structure alignment. Here, we present the main aspects of this measure and we compare it to RMSD-based approaches.

2 Methods

There are various ways to represent three dimensional structure of RNA molecule (Williams and Fleming, 1996; Westhof and Auffinger, 2000). In algebraic represen-

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]Institute of Bioorganic Chemistry, Noskowskiego 12/14, 61-704 Poznan, Poland
[3]E-mail: tzok@skno.cs.put.poznan.pl

tation, being the most common of all, a set of coordinates of all the atoms is engaged to feature the conformation. Second popular representation comes from trigonometric approach. It describes molecule shape by enumeration of dihedral angles. The other means of defining the structure involve inter-atomic and inter-residue distances (geometric representation), electron density distribution (probabilistic representation) or nucleotides and bonds (graph representation). A choice of representation affects the method used to compare structures by the means of distance measure. Root mean square deviation is the most common molecular distance measure. For two structures, *A* and *B*, composed of *N* atoms each, RMSD is calculated due to the following formula:

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(x_{Ai}-x_{Bi})^2 + (y_{Ai}-y_{Bi})^2 + (z_{Ai}-z_{Bi})^2]}\,,$$

where x_{Ai}, y_{Ai}, z_{Ai}, are x, y, z coordinates of the *i*-th atom of *A* and x_{Bi}, y_{Bi}, z_{Bi}, are adequate coordinates of the *i*-th atom of *B*. RMSD depends on reciprocal orientation of structures, thus, it is usually preceded by finding their rigid superposition which minimizes the average distance. Optimal aligmnent of structures can be performed using rotation matrices (Kabsch, 1976) or quaternions (Coutsias et al., 1978). Since finding optimal superposition is quite computationally complex, any idea on how to avoid it seems quite attractive. Therefore, we have proposed a measure independent on molecule position in space. It has been based on trigonometric representation of RNA structure which involves dihedral angles to determine the three dimensional shape.

There are seven dihedral angles in each nucleotide, α, β, δ ε, γ, ζ, χ. First six angles describe the shape of sugar-phosphate backbone, while the final one shows how an organic base bounds to sugar ring. In computation of the distance we use also sugar pucker pseudorotation phase angle P. To find the distance between two conformations represented trigonometrically, we compare the appropriate dihedral angles assuming that the sequences of both structures equal in length. Next, mean of circular quantities (MCQ) is computed basing on the following formulas:

$$\begin{aligned} x = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \quad & [\cos(\alpha_{Ai}-\alpha_{Bi}) + \cos(\beta_{Ai}-\beta_{Bi}) \\ & + \cos(\delta_{Ai}-\delta_{Bi}) + ... + \cos(P_{Ai}+P_{Bi})] \end{aligned}$$

$$\begin{aligned} y = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \quad & [\sin(\alpha_{Ai}-\alpha_{Bi}) + \sin(\beta_{Ai}-\beta_{Bi}) \\ & + \sin(\delta_{Ai}-\delta_{Bi}) + ... + \sin(P_{Ai}+P_{Bi})] \end{aligned}$$

$$MCQ(A,B) = |\arctan(y,x)|$$

where α_{Ai}, β_{Ai}, δ_{Ai} ..., P_{Ai} denote dihedral angles in the *i*-th nucleotide of *A* structure and α_{Bi}, β_{Bi}, δ_{Bi} ..., P_{Bi} – adequate angles in the *i*-th nucleotide of *B*.

Measuring similarity of structures with this approach is less time consuming and easier than in case of RMSD, while the final distance provides at least the same information about the structures. We present an algorithm to calculate dihedral angles

for structures given in pdb format, following by measure of the distances. The results of computational experiment are also given. They show the distances measured by a typical Kabsch algorithm for all the atoms, selected phosphorus atoms and for the sugar-phosphate backbone, as well as these, computed on the basis of dihedral angles.

3 Conclusions

We have proposed the distance measure to compare tertiary structures of RNA molecules. It is based on trigonometric representation of the structure and does not depend on reciprocal orientation of conformations. Since the proposed approach is quite fast it appears as a good tool for structure evaluation, clustering conformations, identifying common patterns within the set of structures, evaluating prediction algorithms, tracing the changes of molecule conformations and aligning small RNA fragments with bigger structures.

4 Acknowledgements

This work was partially supported by grant from the Ministry of Science and Higher Education, Poland.

References

[1] Coutsias, E.A., Seok, C. and Dill, K.A.: Using quaternions to calculate RMSD. Journal of Computational Chemistry, 1978, 25, p. 1849-1857.

[2] Hoehn, L. and Niven, I.: Averages on the move. Math. Mag., 1985, 58, p. 151-156.

[3] Kabsch, W.: A solution for the best rotation to relate two sets of vectors. Acta Crystallographica, 1976, 32, p. 922-923.

[4] Kenney, J.F. and Keeping, E.S.: The standard deviation. Mathematics of Statistics, Princeton, 1962, p. 77-80.

[5] Kavraki, L.E.: Molecular distance measures. 2007, http://cnx.org/content/m11608

[6] Williams, D.H. and Fleming, I.: Spectroscopic Methods in Organic Chemistry. 1996, McGraw-Hill, New York.

[7] Westhof, E. and Auffinger, P.: RNA Tertiary Structure. Encyclopedia of Analytical Chemistry, 2000, John Wiley&Sons, Chichester, UK, p. 5222-5232.

3D–RNA–Pred: An Automatic Construction of Three–Dimensional RNA Structures

Maciej Antczak[2,3], *Jacek Blazewicz*[1,2], *Ryszard Adamiak*[1], *Piotr Lukasiak*[1,2], *Mariusz Popenda*[1], *Marta Szachniuk*[1,2], *Grzegorz Palik*[2]

Acknowledgements: The work has been partially supported by the Polish Ministry of Science and Higher Education grant N N519 314635.

1 Introduction

Ribonucleic acid (RNA) is an important biological molecule. It plays a key role in the synthesis of protein from deoxyribonucleic acid (DNA). It is also known from its structural and catalytic roles in the cell. For the purpose of structure prediction, it can be simply described as a flexible single-stranded biopolymer. The biopolymer is made from a sequence of four different nucleotides: adenine (A), cytosine (C), guanine (G) and uracil (U). Intramolecular base pairs can form between different nucleotides, folding the sequence onto itself [1]. Recent discoveries have demonstrated the role of RNA as biological regulator as well as information-transfer molecule. For example, RNA molecules have been associated with enzymatic functions, gene transcriptional regulation and protein biosynthesis regulation [2].

Knowledge of the 3D structure and dynamics of RNA as well as its interactions with other biomolecules is important for understanding its function in the cell. Experimental techniques such as X-ray crystallography of single crystals of purified RNA molecules, NMR spectroscopy and cryo-electron microscopy used to derive a 3D structure are time consuming and expensive. Only a limited number of attempts have been carried for automatic prediction of the 3D structure of a large RNA molecule. Moreover, the

[1]Institute of Bio-organic Chemistry, Polish Academy of Sciences, Poznan, Poland
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]E-mail: mantczak@cs.put.poznan.pl

complexity and flexibility of RNA molecules makes the determination of 3D structures even more difficult. Thus, the disparity is increasing between known RNA 3D structures and known RNA sequences. This encourages the use of computational methods to obtain information on RNA 3D conformations [3].

2 Problem Formulation

Despite that, the application of computational algorithms of 3D RNA structure prediction has been one of the sources for characterizing the structural diversity in RNA molecules and its relationship to function. Most of the existing algorithms based on the assumption that RNA folding is a hierarchical process and knowledge of its secondary structure may improve the prediction of its 3D conformation. Consequently, several ab-initio methods have been implemented in computational programs for prediction of the base pairs interactions in the RNA from its sequence. However, the growing number of available structural data of RNA molecules and the initial attempts for classification of their motifs have opened a possibility for applying the comparative approaches for RNA structure prediction [3].

In the absence of 3D RNA structures, many processes are envisaged using predicted 2D RNA and experimentally elaborated representations. RNAs and RNA complexes (with proteins or/and small molecules) deposited in the structural databanks (PDB, NDB) are the main source of experimentally proven restraints which together with biochemical/biophysical data might be used in computer-aided modelling of 3D RNA structures and their interactions. In all cases, based on sequence only, 2D RNA prediction is an obligatory step. In most cases such 3D modelling is based on the force field restrained molecular simulations and usually results in the structures of resolution comparable to that given by NMR or X-ray. Nowadays several dedicated software packages for calculations of 3D RNA structures exists, however, all of them lack the speed expected of a major calculation engine that would be capable of the high-throughput prediction of 3D RNA structures [4].

Proposed method tries to predict the tertiary structure of unknown RNA molecule identified only by its sequence based on comparative modelling approach, that has been applied to protein structure prediction for more than two decades.

3 Method

The automatic and intelligent computational algorithm for the tertiary structure prediction based on the primary structure of RNA molecules is proposed.

The idea was to design a publicly available server to compute and analyse data delivered from well proven software dedicated to: 2D structure prediction, generation of the 3D structure, its refinement and analysis.

In general the 3D-RNA-Pred approach consists of the following major stages:

- unknown RNA sequence serves as the input to the most commonly used RNA secondary structure prediction program (for example: MFOLD [5], UNAFOLD [6], RNAfold [7]). As the result *the set of predicted 2D structures* is obtained,
- all 2D structures from *the predicted 2D structures set* are filtered in order to check theirs correctness based on the biochemical/biophysical data. As the result of this stage, *the possible set of 2D structures is obtained*,
- for each 2D structure from *the possible set of 2D structures* the tertiary structure is predicted in according to fulfilled rules defined by experts,
- tertiary structure obtained in previous step is subsequently refined using X-PLOR program [8] to yield the final 3D RNA structure. The refinement is an option which takes more time.

4 Implementation and Tests

Currently tools for major stages of the algorithm are being implemented and analysed from the conceptual and usefulness point of view. One can see that the 3D-RNA-Pred approach is achieving satisfactory solutions and will be a very good choice for those who want to predict the tertiary structure on unknown RNA molecules based only on theirs primary structure.

5 Discussion

The new, useful approach for tertiary structure prediction of the unknown RNA molecules has been proposed. Proposed approach was analyzed from different point of views (for example: efficiency, the prediction quality) and obtained results are proving usefulness but it has to be improved in the future. The publicly available server should be developed in order to combine all constructed tools into one fully functional system useful for 3D structure prediction of unknown RNA molecules.

References

[1] K.C. Wiese, A.A. Deschenes, A.G. Hendriks. RnaPredict-An Evolutionary Algorithm for RNA Secondary Structure Prediction. *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, 5,1:25–40, 2008.

[2] E. Capriotti, Marti–Renom, A. Marc. Computational RNA Structure Prediction. *CURRENT BIOINFORMATICS*, 3:32–45, 2008.

[3] B.A. Shapiro, Y.G. Yingling, W. Kasprzak, E. Bindewald. Bridging the gap in RNA structure prediction. *SCIENCEDIRECT, CURRENT OPINION IN STRUCTURAL BIOLOGY*, 17:157–165, 2007.

[4] M. Popenda, L. Bielecki, R.W. Adamiak. High-throughput method for the prediction of low-resolution, three-dimensional RNA structures. *Nucleic Acids Symposium*, 50:67–68, 2006.

[5] D.H. Mathews, M. Zuker, D.H. Turner. Algorithms and Thermodynamics for RNA Secondary Prediction: A Practical Guide. *Kluwer Academic Publishers*, 1999.

[6] N.R. Markham, M. Zuker. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Research*, 33:577–581, 2005.

[7] N.R. Markham, M. Zuker. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 2005.

[8] A.T. Brunger, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissing, I.N. Shindyaloy, P.E. Bourne XPLOR manual Version 3.1. *NEW HAVEN: YALE UNIVERSITY PRESS*, 1992.

Workers Assignment Simple Assembly Line Balancing Problem - 2

Tomasz Kujawa[1,3]***, Janusz Lembicz***[1]***, Grzegorz Pawlak***[1]***, Alf Kimms***[2]

1 Introduction

In the paper the production scheduling problem drown from the car factory was considered. Assembly line balancing problems are important tasks in production planning. The original car assembly line problem was formulated as a permutation flow shop problem in the multi-stage system. The classification of Assembly Line Balancing Problems was introduced in [2].

A number of attempts has been made to develop mathematical formulations of Simple Assembly Line Balancing Problems (SALBP). The first formulation of SALBP was described by Salveson [1] - this model requires the pre-determination of all possible station loads, so it is hard to apply the approach in the practice. SALBP-2 i characterized by the minimization the cycle time for a given number of stations and was presented in [2]. Also this approach is not a rule of thumb.

In the paper the problem was extended by introducing workers with skills. The skills determine the set of tasks the workers can process. It makes the problem even more difficult but more appropriate to the real situations.

The purpose of the paper is to minimize the cycle time which in this case is equivalent to minimize the flow time of the cars in the assembly line and the idle time of the workers at the stations. The presented problem is shown to be NP-hard. For the problem described above the mathematical model and algorithms have been designed. The motivation for the research was drown from the real car factory.

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]University of Duisburg-Essen, Germany
[3]E-mail: tkujawa@skno.cs.put.poznan.pl

2 Problem formulation

For each car there are certain requirements represented as a precedence graph $G = (V, A, t)$ which is a non-cyclical digraph with a set $V = \{1, \ldots, n\}$ of n nodes and a set $A = \{(i, j) : i \in V \wedge j \in V\}$ of arcs. Tasks are represented as nodes and direct precedence relations between tasks are represent as arcs. As an input there are also sets of w workers and m stations.

The goal is to assign the operations to stages in the way that the cycle time will be minimal, with respect to the precedence constraints.

3 Solution algorithms

3.1 Branch and Bound Algorithm

A Task Oriented Branch and Bound Procedure (TBB) proposed in [2], is not feasible in the SALBP-2 with Workers Assignment problem. TBB-2 does not use information about workers and assumes that tasks can be done on every station in series way.

SALBP-2 with Workers Assignment problem widens the problem structure and for that specific problem a branch and bound procedure was proposed. Main parts of the algorithm are as follows:

Preprocessing: compute *LBs* and *UBs*

1. Set $c = min\{UB\}$
2. Check if the assignment for c is possible
3. **YES**: $C = c$ and decrease c and go back to the point 2

   NO: stop the algorithm and show results
4. Minimal cycle time is stored in C

The algorithm which checks the feasibility of the line with specified cycle time uses several components such as Lower Bounds, Upper Bounds, Earliest and Latest Stations.

Lower and Upper Bounds Basically, Upper Bounds may be obtained from a theoretical upper bound or any heuristic algorithm. Since theoretical bounds are usually rather weak it is better to utilize a heuristic procedure as one presented in this paper.

For SALBP-2 with Workers Assignment were proposed several Lower Bounds. LB were used to define the range of cycle time.

Earliest and Latest Stations. For each task j the earliest stations $E_j(\overline{c})$ and the latest stations $L_j(\overline{c})$ based on the cycle time $\overline{c}$ were defined. The station interval is defined as follows:

$$SI_j(\overline{c}) \; = \; \{E_j(\overline{c}),\; E_j(\overline{c}) + 1, \ldots,\; L_j(\overline{c}) - 1,\; L_j(\overline{c})\} \tag{1}$$

The way of assigning components of the problem causes that the temporary station interval must be defined:

$$tSI_j(\overline{c}) \;=\; \{k : k \in SI_j(\overline{c}) \;\wedge\; Zj \in w[SI_j(\overline{c})]\} \tag{2}$$

where $w(S_k)$ means the set of tasks possible to be done on the station k.

3.2 Heuristic Algorithm

Since SALBP-2 is considered NP-hard also SALBP-2 with Workers Assignment which is its extension is also NP-hard. It is fully justified to develop a heuristic solving method in order to achieve good results in a reasonable computational time.

The heuristic approach was mainly focused on making tasks as simultaneously as possible. The algorithm consists of 5 basic steps:

1. Building sets of potentially parallel tasks (layers) Disjunctive sets of tasks which can be make separately are constructed.

2. First estimation workers to layers assignment.

   Greedy assignment workers to layers provide first execution time estimation.

3. Deploying layers on stations – dynamic programming approach.

   Dynamic programming algorithm have been designed to deploy layers on stations minimizing cycle time difference between them in polynomial time. After this step feasible solution is constructed.

4. Optimization of workers assignment.

Result of previous steps could be optimized by performing described algorithm on each station.

4 Conclusion

The problem of finding the shortest cycle time on the assembly line was defined and algorithms were proposed. The future work will be considered on finding better lower and upper bounds. Better browsing way through the range of possible time must be also enhanced.

The heuristic algorithm presented in this paper could be improved by some local search algorithms such as Tabu Search. With already gathered information it is easy to plan potentially attractive moves considering layers of parallel tasks.

References

[1] Salveson M. E., *Assembly Line Balancing Problem*, Journal of Industrial Engineering, 6 (1955), pp. 18-25.

[2] Scholl A., *Balancing and sequencing of assembly lines*, Physica Verlag, Heidelberg 1999, pp. 54-60.

Vehicle Scheduling in the Car Factory Paint Shop

Grzegorz Pawlak[1,2]***, Marek Rucinski***[1]

1 Introduction

In the modern car factory the problem of increasing productivity is always important. In the paper the problem of scheduling cars in the paint shop of the car factory has been considered. The particular problem is modelled and simple algorithms, for the flow control of the cars through the production line of the paint shop, have been proposed.

The problem is to schedule cars in order to increase the throughput rate through the painting station of the base color painting machine stage. The throughput rate depends on the size of the car blocks of the same colors and the car types in the car sequence constructed by the production plan. The input cars sequence is optimized in the input buffer where the car bodies are sorted in order to minimize the color switchings in the painting machine stage. Then cars are painted in the painting station and then go to the inspection station where the quality of the painting is checked and some of them are redirected back to the input sequence for repainting. The previous input sequence is interrupted by the reentrant cars and the number of color changes in the painting station increases. According to the technological process the repainting of the same car may be done several times. Some of the cars are obligatory repainted because of the double base color layers demanded. The reentrant line consists of buffer where the cars could be sorted. The repainting of the care has time window for the introduction to the original car sequence limitation and the processing must avoid deadlock in the process of painting line.

The Figure 1 shows the production area and the process chart. The BC represents the base color painting machine; $B1, B2, B3$ represents limited input, reentrant line and output buffers, respectively; IS states the /it Inspection Stage where the car routing is decided whether car goes to the output buffer or to the repaint line and buffer $B2$.

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]E-mail: grzegorz.pawlak@cs.put.poznan.pl

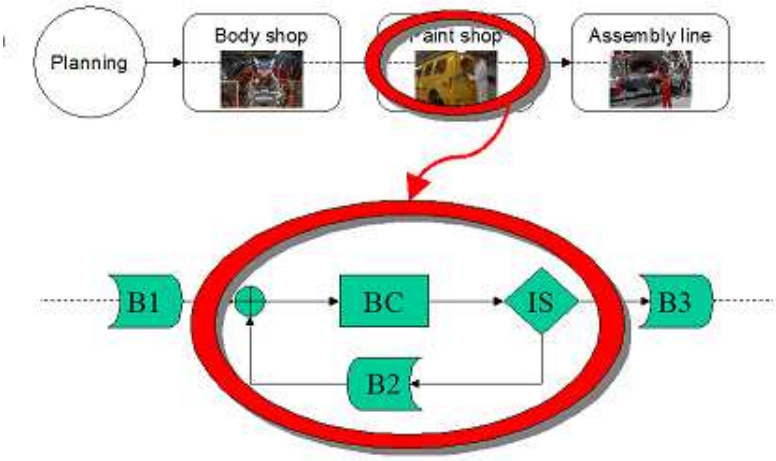


Figure 7: The production area chart

2 The problem formulation

The problem has been formulated as the single machine scheduling problem with reentrant jobs and set-up times $J_j = (T_j^1, T_j^2, \ldots, T_j^{n_j})$ and $n_j \geq 1$ where n_j is the number of painting operations for each car. Each job has addition two properties $type_j$ representing the type of the car and the $color_j$ - the color of the car. The processing time p_j is represented by the time necessary to paint the car of the particular type. Between the subsequent cars the set-up time is introduced. There are two types of the set-ups. First s_i introduced when two subsequent cars have the same color and second s_j when two subsequent cars have different colors. The set-up time is connected with the cleaning process of the painting guns. The set-ups holds inequality $0 \leq s_i < s_j$. On the re-paining line the buffer $B2$ is located with finite capacity N_B. The cars must leave buffer $B2$ within the /it Time Window $T_j^i : (r_{T_j^i}, d_{T_j^i})$ where $r_{T_j^i}$ is the ready time of the car to be repainted and $d_{T_j^i}$ is the deadline when the car must be introduced into the input sequence of the cars and painted at the painting station. The input car sequence is predefined by the production plan. The criterion is C_{max}. Because of the fixed speed of the line the C_{max} criterion is equivalent to the throughput rate of the production in the defined time horizon. In addition the number of the external set-up times has been also measured.

Figure 8: Gain for applying the proposed algorithm in comparison with the existing control algorithm

3 Complexity of the problem

The deterministic version of the problem with n_j known in advance is in the general case NP-hard. One can easy transform NP-hard scheduling problem denoted as $1|chain(l)|C_{max}$, described in [1], into the described above. The presented problem is even more complex by introducing the set-up times, time windows and finite capacity buffer. Moreover, in the real case the value of n_j is not known in advance. Then the on-line approach could be applied.

4 Algorithms and results

In the paper two on-line algorithms were presented. The algorithms were tested on the real data. The data were collected from the car factory in the one month time horizon. The algorithm were compared to the existing algorithm controlling the process. The original one were based on the FIFO rule. The proposed algorithms utilized the different dispatching rule. In the given time window the cars were sorted in the buffer and introduced to the main sequence maximizing the car block size of the same color. The result were presented in the Figure 2. The gain in comparison to the number of job colors. Introducing the proposed algorithms can gain for increasing number of the job colors from 15 to 25.

5 Conclusions

In the paper the practical approach for the solving the real car scheduling problem were proposed. The car painting problem was modelled as the one machine scheduling problem with the reentrant jobs. The proposed on-line algorithms were tested on the real data and their effectiveness was showed. Further research will consider meta heuristic approach for solving the problem more accurately.

References

[1] E.D. Wikum, D.C. Llewellyn, and G.L. Nemhauser, *One-machine generalized precedence constrained scheduling problems* Operations Research Letters, 16(2), 1994, pp. 87-99.

Basic Concepts of Quantum Computing

***Wojciech Mruczkiewicz*[1], *Hanna Cwiek*[1,3], *Radoslaw Urbaniak*[1], *Piotr Formanowicz*[1,2]**

1 Introduction

Even though the concept of a quantum computer was first conceived in the 1980's, today it is still in its infancy, as practical realisation of the idea proves a number of difficulties to scientists. The theory, however, develops intensively and a number of publications are published. It is discussed whether the possibilities offered by quantum mechanics can help us solve difficult problems effectively. For the time being, a few types of problems have been found whereby quantum computing proves useful.

We will introduce some basic aspects of quantum computation. The topic was presented by numerous authors, e.g. [6]. We will also show two quantum algorithms to give a general idea about what could be expected.

2 Basics

The major concept in quantum computing is a qubit. A qubit is an equivalent of a classical bit – it is a quantum system whereby classical boolean values 0 and 1 are represented by a pair of quantum states, denoted by $|0\rangle$ and $|1\rangle$. The states are mutually orthogonal and normalised, and they form a computational basis. Qubit's state is a linear combination of the base states $a|0\rangle + b|1\rangle$. The complex coefficients a and b are called amplitudes of the base states. They satisfy the condition $a^2 + b^2 = 1$.

Qubits capability of being in two states simultineously has a specific property – it cannot be observed directly. Superposition exists only as long as the quantum system is isolated from environment and no measurement is taken. When measured, a qubit settles its value so that the result of measurement is always a pure state. The result is

[1]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
[3]E-mail: hcwiek@pandora.cs.put.poznan.pl

random, yet the probability of measuring each state is described by its squared amplitude – a^2 and b^2. After the first measurement the state of a qubit is known for certain and all further measurement are bound to give the same result.

A collection of qubits is called a quantum register. Due to superposition, an n-qubit register can store up to 2^n states simultineously. In formal notation, quantum register can be represented as $|a_{n-1}\rangle \dots |a_1\rangle|a_0\rangle$.

Qubits' capability of being in two states simultineously is used in quantum computations. The computations need to be reversible, which means that no information can be lost. Computations are performed by changing qubits' states by quantum gates. Single operations can be presented as applying operators to quantum registers.

There are two types of quantum gates – traditional-like and typically quantum ones. The first group consists of gates being equivalents of traditional logic reversible gates. The examples might be cNOT or Toffoli gate. Typically quantum gates have no classical equivalents and include such gates as $\sqrt{NOT}$, phase shift or Hadamard-Walsh. Hadamard-Walsh gate is a single qubit gate whose action is known as Hadamard transform – each base input state is turned into superposition of the base states. If the input for any quantum gate consists of qubits in the superposition all the base states are processed simultineously. If the 'result set' contains the same base states their amplitudes are summed up, which results in increasing or decreasing the amount of the base state in the output superposition.

3 Quantum algorithms

Constructing a quantum algorithm means finding a sequence of qubit operations that performs computation on all possible states simultineously and manipulates the amplitudes in a way that leads to reading the correct answer for the problem.

One of the first quantum algorithms was presented by David Deutsch [5][3]. In Deutsch problem there is given a function $f : \{0,1\} \to \{0,1\}$ operating on boolean values, whose complete definition is unknown. The task is to determine whether f is constant or balanced, i.e. whether the values of f for the two possible different arguments are equal or not. Solving the problem on a classical computer requires two evaluations of function f and comparing the outputs. A quantum algorithm requires only one evaluation of f to solve the problem.

Quantum circuit used in Deutsch algorithm is presented in Figure 9. The input is a two-qubit register $|x\rangle|y\rangle$ set to value $|0\rangle|1\rangle$. H denotes Hadamard gate. U_f is a specially prepared quantum gate that takes two qubits, $|x\rangle$ and $|y\rangle$, and adds modulo two the value of $f(x)$ to the value of y, i.e. $U_f : |x\rangle|y\rangle \to |x\rangle|y \oplus f(x)\rangle$. After the transformations the final state of the register is

$$\frac{1}{2}\left[\left((-1)^{f(0)} + (-1)^{f(1)}\right)|0\rangle + \left((-1)^{f(0)} - (-1)^{f(1)}\right)|1\rangle\right]\frac{1}{\sqrt{2}}\left(|0\rangle - |1\rangle\right) .$$

Measuring the first qubit gives the answer to the problem – the result is $|0\rangle$ only if f is constant, i.e. $f(0) = f(1)$, and $|1\rangle$ otherwise.

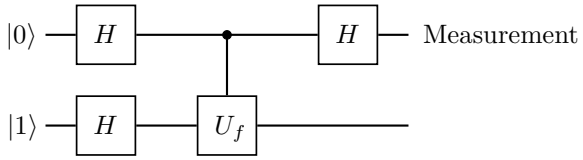


Figure 9: Quantum circuit for Deutsch algorithm.

Deutsch algorithm began the search for other quantum algorithms. Lov Grover presented one that could be used to solve a wide class of problems [7]. In Grover Search a function $f_y : \{0,1\}^n \rightarrow \{0,1\}$ is given. The algorithm is to find such a value of x for which $f_y(x) = 1$. The condition is met only for certain value y. The task is quite a general one, as the function f_y can encode many well known problems, e.g. the SAT problem [4], in which case the bits of the register x are mapped to boolean variables and the value of the function f_y is equal to 1 when the assignment is satisfied, and 0 otherwise.

The algorithm starts with a quantum register in a superposition state. The initial state is transformed by a series of Grover algorithm iterations. In every iteration two quantum gates are applied successively and the probability of measuring the register $|x\rangle$ in state $|y\rangle$ is slightly increased while the probabilities of reading other states are decreased. Grover proved that it is required to perform $\Theta(\sqrt{2^n})$ steps to get a correct answer with probability close to 1. Classical counterpart of Grover algorithm is an extensive search through entire domain of f_y. When nothing is known about f_y the complexity of this search is $\Theta(2^n)$. Grover search improved classical algorithm by a quadratic factor. Benett et al. [1] proved that the algorithm is optimal on quantum computers. No quantum algorithm can search a database faster than $\Theta(\sqrt{2^n})$ without exploring the intrinsic structure of function f_y.

4 Summary

Presented algorithms show certain speedup of quantum computers over classical ones. Moreover, there exist quantum algorithms that, relative to oracle, give an exponential speedup[2] over known classical algorithms. However, the question of how powerful quantum computers really are and whether they could solve NP-complete problems in polynomial time remains unanswered.

References

[1] Charles H. Bennett, Ethan Bernstein, Gilles Brassard, and Umesh Vazirani. Strengths and weaknesses of quantum computing. *SIAM J. Comput.*, 26(5):1510–1523, 1997.

[2] Gilles Brassard and Peter Hoyer. An exact quantum polynomial-time algorithm for simon's problem, 1997.

[3] Richard Cleve, Artur Ekert, Chiara Macchiavello, and Michele Mosca. Quantum algorithms revisited, 1997.

[4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Second Edition*. The MIT Press, September 2001.

[5] David Deutsch. Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London Ser. A*, A400:97–117, 1985.

[6] Artur Ekert, Patrick Hayden, and Hitoshi Inamori. Basic concepts in quantum computation, 2000.

[7] Lov K. Grover. A fast quantum mechanical algorithm for database search, 1996.

From Documents Processing to an Identification of Marine Organisms' Habitat Specificity

Marta Szachniuk[1,2,3]***, Marcin Radom***[1]***, Agnieszka Rybarczyk***[1]***,***

Piotr Formanowicz[1,2]***, Jacek Blazewicz***[1,2]

1 Introduction

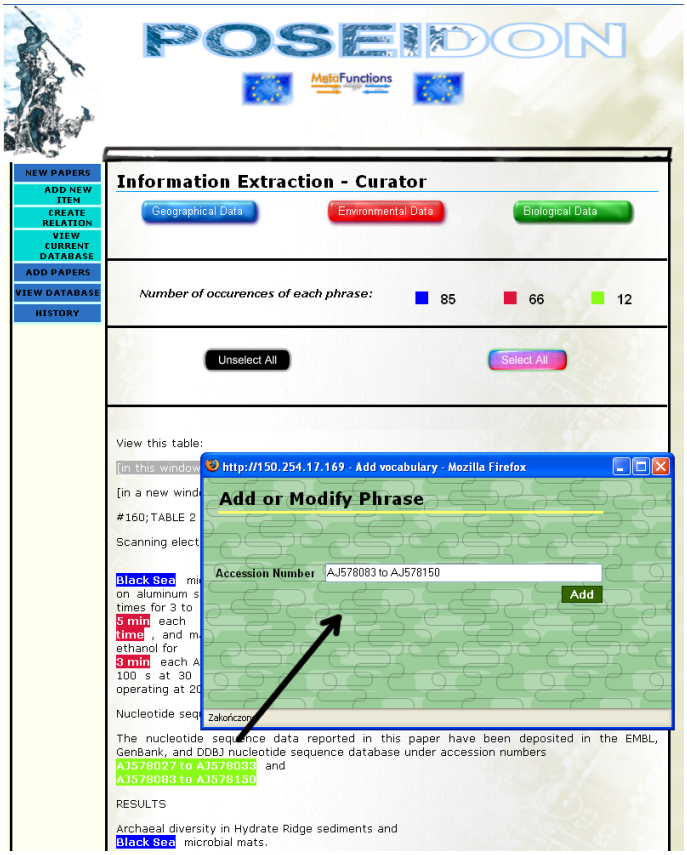
Metafunctions [1] is a research project funded by the European Comission within the 6th Framework Programme under the NEST – Newly Emerging Science and Technology Adventure initiative. This three-year project (2005-2008) assembles four teams, from three European countries, which bring expertise in marine microbiology, biotechnology, molecular ecology, biogeochemistry and computer science. The project aims at creating a bioinformatic system to detect and assign functions to habitat specific gene patterns. It is dedicated to an analysis of metagenomes collected from marine environments.

2 Description

More than 99% of the microbial diversity on earth still resists cultivation. To address their metabolic potential, numerous efforts to clone and sequence large DNA-fragments directly from the environment (the metagenome) have been started worldwide. Unfortunately, still more than 50% of the genes found on sequenced genomes and metagenomes lack functional assignments. Metafunctions project addresses the issue of combining genes with functions by proposing a three stage process consisting of (i) correlation of available genomic information with geographical, geological, biological and physical/chemical data, (ii) identification of habitat specific gene patterns, (iii) prediction of

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]Institute of Bioorganic Chemistry, Noskowskiego 12/14, 61-704 Poznan, Poland
[3]E-mail: marta.szachniuk@cs.put.poznan.pl

functions for the identified patterns of conserved genes. Concluding on habitat specificity of genes requires an analysis of huge amounts of data which can be gathered during sampling expeditions. However, a lot of valuable information about the results concerning marine microorganisms is also found in a variety of scientific papers. Thus, one of the project tasks was the development of a unique system for a supervised analysis of marine metagenomic papers. Such an analysis includes paper classification, information retrieval and extraction, as well as supplying the database with newly collected data. All of these procedures define the functionality of Poseidon system that has been designed and implemented by the team of Poznan University of Technology. Poseidon maintains scientific papers collected from a specified list of journals. The temporary repository of papers is automatically updated once a month. Next, papers are processed in a pipeline by several independent modules of the system. At first, each paper is analyzed by a conversion and normalization module (Augeas) which adapts the document into the format easy to handle by other tools of Poseidon. Next, the paper goes through the procedures performing an initial analysis of its contents (Cerberus module). This step reduces the number of documents accepted for further processing. Only the papers concerning metagenomics and marine environment are considered relevant and accepted for further analysis. These papers are passed along to Whatizit module [2] (a tool prepared in European Bioinformatic Institute), which processes the text using specified keyword dictionaries (i.e. bacterial names, geographical names and others). Keywords and phrases specified in the dictionaries are tagged within the document's body. Finally, the text with tagged phrases is shown via Trident module to the curator, who decides either that the extracted data are correct and can be placed in the database or that they require some verification (completion, modification). The Poseidon interface provides the possibility to verify and modify the extracted entities. Then accepted by the curator, the extracted information is sent to the database. Figure below shows text of selected paper with some highlighted keywords, among them there is an accession number (also presented in a separate window which opens when the curator clicks onto the keyword).



3 Summary

We have presented Poseidon – a unique system for classification of scientific documents related to marine metagenomics and extraction of biological and environmental data. The information extracted from the documents is processed under the curator supervision and it supplies MetaStorage database as well as Genome MapServer [3] and MetaMine – being the other tools prepared for the purposes of the project.

References

[1] MetaFunctions: http://www.metafunctions.org

[2] Whatizit: http://www.ebi.ac.uk/webservices/whatizit

[3] Genome MapServer: http://metafunctions.grid.unep.ch/mapserver/

TU Clausthal
ICOLE 2008, LESSACH, AUSTRIA

GeVaDSs - A System for New Improved Vaccines Based on Genomic and Proteomic Information

Piotr Lukasiak[1,2,4]***, Jacek Blazewicz***[1,2]***, David Klatzmann***[3]

CompuVac [1] is a project financed by the European Commission, which involved 18 partners worldwide. CompuVac's main objectives are to setup a standardized approach for the rational development of genetic vaccines and to apply this methodology to the development of vaccines against the hepatitis C virus. Main objectives of CompuVac are:

- to standardize the qualitative and quantitative evaluation of genetic vaccines using defined *"gold standard" antigens and methods*
- to rationally develop a *platform of novel genetic vaccines* using genomic and proteomic information, together with our gold standards
- to generate and make available to the scientific community a *"tool box"* and an *"interactive database"* allowing to comparatively assess future vaccines to be developed with our gold standards

The process comprises the development of: (i) a large panel of vaccine vectors representing various vector platforms and all expressing the same model antigens; (ii) standardized methodologies for the evaluation of T- and B-cell responses and of molecular signatures relevant to safety and efficacy; (iii) a database for data storage and analysis of large data sets; (iv) intelligent algorithms for the rational development of prime boost vaccination. One of our main goals is to generate and make available to the scientific community a "tool box" and an "interactive database" allowing the comparative

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland
[3]Universite Pierre et Marie Curie, 7087 - UPMC-CNRS, Paris, France
[4]E-mail: `Piotr.Lukasiak@cs.put.poznan.pl`

assessment of future vaccines. We also aim to validate these tools by the rationale development of preventive and/or therapeutic vaccines against HCV. We have now assembled a unique set of 142 vaccines of different class, from viral vector derived vaccines to inert VLPs, analyzed their efficacy with standardized methodologies, and compared them with an intelligent database. This has already allowed us to make significant comparisons between different vaccine types and to initiate novel vaccine design and vaccination regimen. We are now evaluating prime-boost immunization regimen with these vectors. We believe that this should have significant impact on vaccine development, and notably for those vaccines requiring prime/boost immunizations. (Figure 10 on page 68.)

"Gold standard" algorithms for intelligent interpretation of vaccine efficacy and safety will be built into Compuvac's interactive "Genetic Vaccine Decision Support System", which should generate (i) vector classification according to induced immune response quality, accounting for gender and age, (ii) vector combination counsel for prime-boost immunizations, and (iii) vector safety profile according to genomic analysis. The consortium will generate a *toolbox* and an interactive database termed GeVaDSs (Genetic Vaccine Decision Support system) [2]. The toolbox will contain formatted data related to our defined *gold standard antigens and methods* used to assess immune responses. The interactive database will contain formatted data related to results obtained using our gold standard antigens and methods, i.e. newly acquired results as well as pre-existing results retrieved through data mining, and *algorithms* allowing the intelligent comparison of new vectors to previously analyzed ones. At this stage, a large panel of vaccine vectors has been produced. It is important to note that many different improvements have been made to our vectors, some vectors design were abandoned or refocused, and some additional vectors were produced. This is a unique asset for vaccine development, since it will allow for the first time a meaningful comparison of these vaccines, as single immunogens and in association. To this aim, we have already successfully developed and standardized methods using GeVaDSs to measure the efficacy and safety of individual vaccine vectors, in a manner that allows comparison between different vaccine designs, tested in different laboratories, at different time points. With these methods, we have already analyzed the efficacy of a unique set of vaccines, and compared them with an intelligent database. GeVaDSs has allowed us to make some significant comparisons between different types and to initiate novel vaccine design and vaccination regimens. Besides monitoring of T- and B-cell immune responses, we also aimed at monitoring vaccine "efficacy" and "safety" profiles by analyzing relevant molecular signatures obtained from transcriptomes studies. The "efficacy" profile has now been validated and the "safety profile" is still being developed, based on analyzing molecular signatures from whole liver and spleen after injection of vaccine vectors. The results of these experiments will drive the development of our HCV vaccines. We have already tested the first HCV vectors generated in single immunization regimen, and obtained interesting results suggesting the great potential for the association of our two classes of vectors, viral and VLP derived. CompuVac aims at making GeVaDS system accessible to any researcher developing genetic vaccines. Retrieving previously gener-

ated or introducing newly acquired results obtained with validated approved methods, should allow any researcher to rationally design and improve his or her vaccine vector as well as to comparatively assess its efficacy and potential.

References

[1] CompuVac: `http://www.compuvac.org`

[2] GeVaDS: `http://gevads.cs.put.poznan.pl`

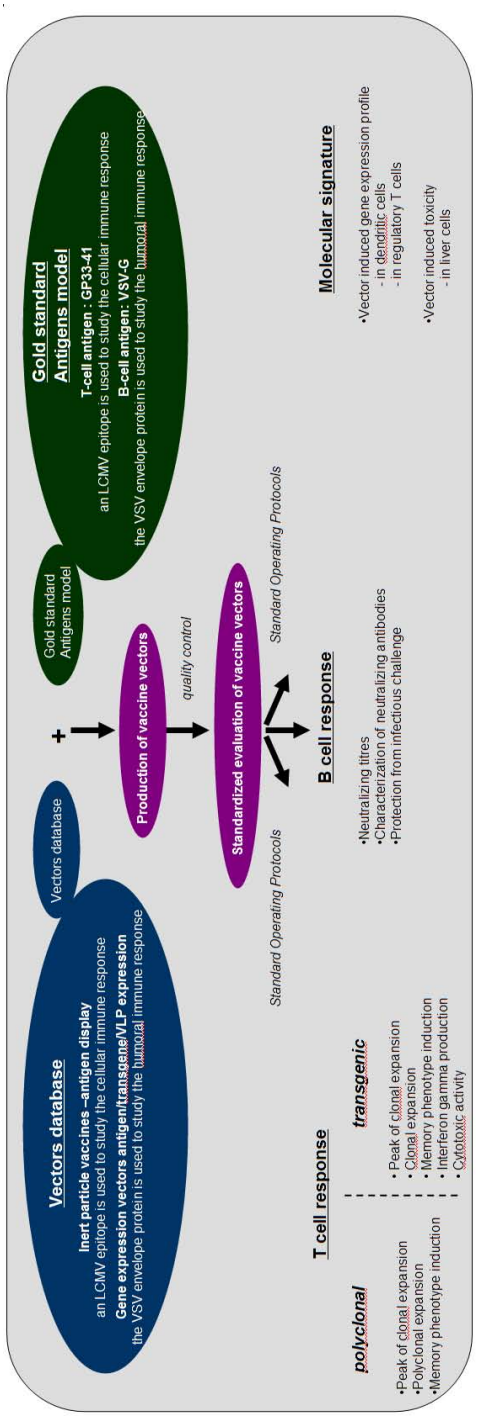


Figure 10: Visualization of the process of development of novel genetic vaccines

TU Clausthal
ICOLE 2008, LESSACH, AUSTRIA

Scheduling Problem Applicable to 'Simulation of Crisis Management Activities' (SICMA) EU Project

Michal Tanas[1,2,3,4]***, Witold Holubowicz***[1,2]***, Rafal Renk***[1,2]

Acknowledgement: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 217855 (SICMA Project).

1 Introduction

A scheduling problem is, in general, a problem answering a question of how to allocate some resources over time in order to perform a given set of tasks [1]. In practical applications resources are processors, money, manpower, tools, etc. Tasks can be described by a wide range of parameters, like ready times, due dates, relative urgency factors, precedence constraints and many more. Different criteria can be applied to measure the quality of a schedule.

Scheduling theory is widely applicable to solve real life discrete optimization problems. The purpose of this paper is to present another application of particular scheduling problems to the „Simulation of Crisis Management Activities" (in brief SICMA) EU project, whose objective is to improve health service crisis managers decision-making capabilities through an integrated suite of modelling and analysis tools providing insights into the collective behavior of the whole organization in response to crisis scenarios

[1]Applied Computer Science Division, Physics Faculty, Adam Mickiewicz University, Poznan, Poland
[2]ITTI sp. z.o.o., Poznan, Poland
[3]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[4]E-mail: Michal.Tanas@cs.put.poznan.pl

2 Description of the SICMA project

SICMA is EU FW7 Security Research Call 1 programme, whose point of interest is how to restore security and safety after a crisis. The term „crisis" denotes any unexpected and harmful event from the level of serious local accident (e.g. a car crash involving a truck carrying dangerous chemicals) to the level of state wide disaster (e.g. a large scale leak in a chemical factory, like Bhopal disaster in 1984). In order to enhance the efficiency of command and control by intelligent decision support systems, the task of the SICMA project is to develop appropriate novel approaches to computer assisted decision making. Applications should be robust and facilitate the cooperation of operational units across organizational boundaries. Two main research goals of the project are as follows:

1. The first challenge is to ensure that governments, first responders and societies are better prepared prior to unpredictable catastrophic incidents using new, innovative and affordable solutions.

2. The second challenge is to improve the tools, infrastructures, procedures and organizational frameworks to respond and recover more efficiently and effectively both during, and after, an incident.

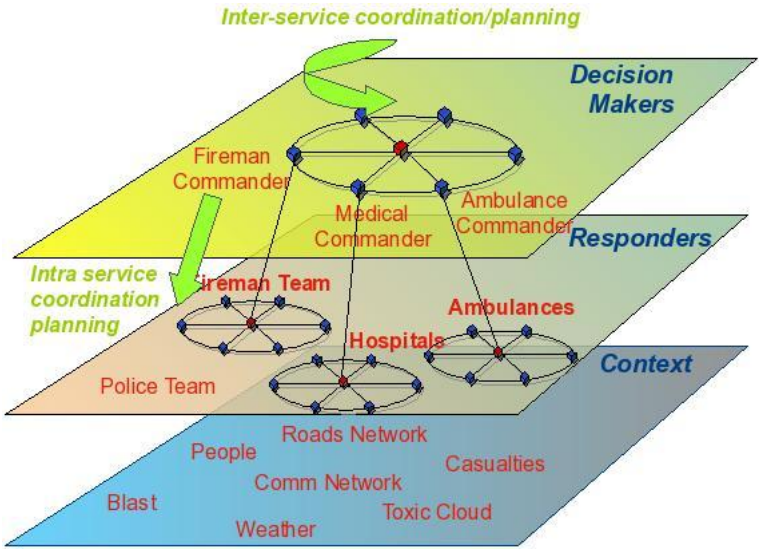
Decision-making support will be provided through an integrated suite of modelling and analysis tools. Key research aspects are:

- „bottom-up" modelling approach
- build independent model components and then combine them
- unpredictable factors modelling
- human behavior, mass behavior
- procedure support
- provide the user with the correct procedures to solve the problem
- computation of the 'distribution' of the effectiveness of a certain 'decision' rather than the effectiveness of that
- solution deterministically dependant on the preconceived scenario The combined effects of the above points will allow to document both the unexpected bad and good things in the organization(s) thus leading to better responses, fewer unintended consequences and greater consensus on important decisions.

3 Structure of crisis management services

In general the response to the crisis is the result of the activities of:

- different services (e.g. police, medical care, rescue forces, fire fighting, etc)
- interacting vertically (i.e. with components of the same organization) and horizontally (i.e. with components of other organizations)
- in a complex environment



4 Example crisis scenario

In context of the SICMA project many different crisis scenarios are considered. One of them is a scenario of industry accident or terrorist attack which results in a conventional explosion which releases a chemical agent and creates a large toxic cloud of chemicals inside a massively urbanized area. In this scenario the whole area of the crisis should be isolated, all unaffected people should be evacuated to create a safety buffer zone, and all victims (i.e. people injured by a explosion or toxic vapors) must be triaged and delivered to neighboring hospitals as fast as possible. The key problems in this scenario are:

- The largest hospitals are capable to take only 7 seriously injured patients per hour without risking of deteriorating quality of treatment. This ratio is surprisingly low.

- Victims must be delivered to hospitals by a transport system (i.e. ambulances) through very dynamic an unpredictable environment (i.e. unpredictable movement of chemical cloud, traffic jams, unpredictable crowd behavior, fires, building collapses, etc.)
- Each hospital has its own „menu" of treatments which it can perform. There are basic procedures which can be performed by any hospital, but there are some highly sophisticated medical procedures (i.e. decontamination of victims of not commonly used chemical agents) which can be performed only by a few hospitals in a state.

5 Application of scheduling theory

Considering all the facts above it is obvious that scheduling theory can be applied to reach the goals stated by EU. In particular in the example scenario presented above there are clear transformation from real-live crisis management problem into a scheduling problem, as follows:

- A hospital can be modeled by a processor
- The number of patients which a hospital can take without deteriorating of quality of service can be modeled by relative speed factor of processors.
- The „menu" of treatments a hospital can perform can be modeled by semi-specialized processors, or by unrelated processors if infinite processing times are allowed.
- A patient can be modeled by a task
- An evacuation of a hospitals (i.e. a hospital is on the way of the cloud) can be modeled by a processors breakdowns
- Triage and search and rescue teams activities can be modeled by a ready time
- Deteriorating of victims health in time can be modeled by due-dates
- Deaths caused by too late help can be also modelled by due-dates (this time with higher penalty for late tasks)
- Ambulance service can be modeled by a transport system. Transport time is variable, there are various amount of different transport vehicles (e.g. ambulances, helicopters)
- Tasks are independent. Treatment of a victim does not depend in any way on treatment of another victim.

- Processing times are equal and tasks are identical. The project does not distinguish details like duration of particular medical procedures. Moreover, a victim which was properly inserted into a hospital is no more object of interest of the SICMA (it is assumed that a hospital knows what to do).

So, a real-life problem considered in the SICMA can be transformed to an unusual unrelated processors problem

$$Q/R|p_j = 1, r_j, d_j, \text{manual intervention}, \text{transport}, \text{machine unavailability}, \\ \text{exclusion lists}, \text{non linear due date penalties}|L_{max}$$

Note, that in spite of the existence of *„exclusion lists"* which for each processor define set of tasks which cannot be processed on this particular processor, and so processors are not fully identical, the problem cannot be considered as job shop, because of the following key facts:

- Each job contains exactly one operation.
- There are several machines which can process a particular job.

6 Conclusion

Among many others practical applications, scheduling theory may be successfully used in health care decision support systems and in modeling crisis management activities, ambulance service in real urbanized areas environment. More over the SICMA EUs project creates several new detailed problems in domain of scheduling theory, which needs to be solved.

References

[1] K. Baker, *Introduction to Sequencing and Scheduling*, J. Wiley, New York, 1974